**海光AI路线金融行业应用优势**

海光信息技术股份有限公司是一家高性能处理器芯片设计企业，主营业务是通用处理器CPU和异构计算处理器DCU（Deep Computing Unit ）的研发、生产和销售。公司以市场为导向，以客户需求为目标，结合自身技术和产品的优势及特点开展研发工作，通过不断优化组织结构和研发管理流程来保证技术创新的持续性。

从2019年开始，研发人员、研发投入、营业收入呈逐年递增趋势。目前正式员工已有XXX余人，研发人员占比高达XX%，研发人员学历以博士、硕士为主。海光CPU和DCU技术来源于AMD，获得指令集层面的授权，拿到了全部的原始代码和微架构，经过最开始的引进、消化到现在吸收、再创新，海光CPU已经发布四代产品，正在研发第五代产品，规划第六代产品。DCU发布两代产品，研发第三代、第四代产品，完全具有产品的自主迭代能力，而且能够确保供应链没有问题。海光用事实证明了有能力实现产品的完全自主可控，不依赖也不受国外技术的限制。

**一、海光DCU在金融行业发展迅猛**

海光DCU于2014年引入我国，主要用于科学计算、智算中心等行业，在2022年9月正式启动民用市场推广，在金融、互联网、运营商、党政、央国企等行业。 2024年全行业年度销售量已超XX万片。

海光DCUK100-AI（深算2号） 于2024年6月份正式发售。目前已中标工商银行总行、中国银行总行、中国邮储银行总行、中国银联、光大银行、华夏银行、中信银行、渤海银行。 已落地千卡GPU集群项目，PICC人保集团944片。海光DCU已成为金融数据中心中标数量最多的国产GPU卡，与华为场景互补（华为在OCR、ASR、声纹类场景难以适配）。

海光DCU 主力卡K100-AI 目标替代英伟达V100、L20、A10。 下一代BW系列性能超越英伟达A100卡。

**1.1 PICC人保项目**

海光DCU 于2024年3月份中标PICC人保国产GPU卡采购项目，DCU作为唯一配件指定写入招标文件。

1、场景需求：

【车险理赔业务】PICC人保财险理赔业务对于OCR系统的支撑有强算力支撑需求。不低于V100。

【司南智能助理】大模型办公软件系统升级应用算力支撑需求。通义千问、百川、SD生图模型等。

【社交媒体服务平台】需要满足目前大模型 ai绘画图片生成运行环境，不低于V100。

【社保开发处】社保资料电子化，医疗票据定制化解析场景，不低于T4。

【大数据应用平台】大数据分析系统，需要满足对yolov8、fast t5模型的支持，不低于T4。

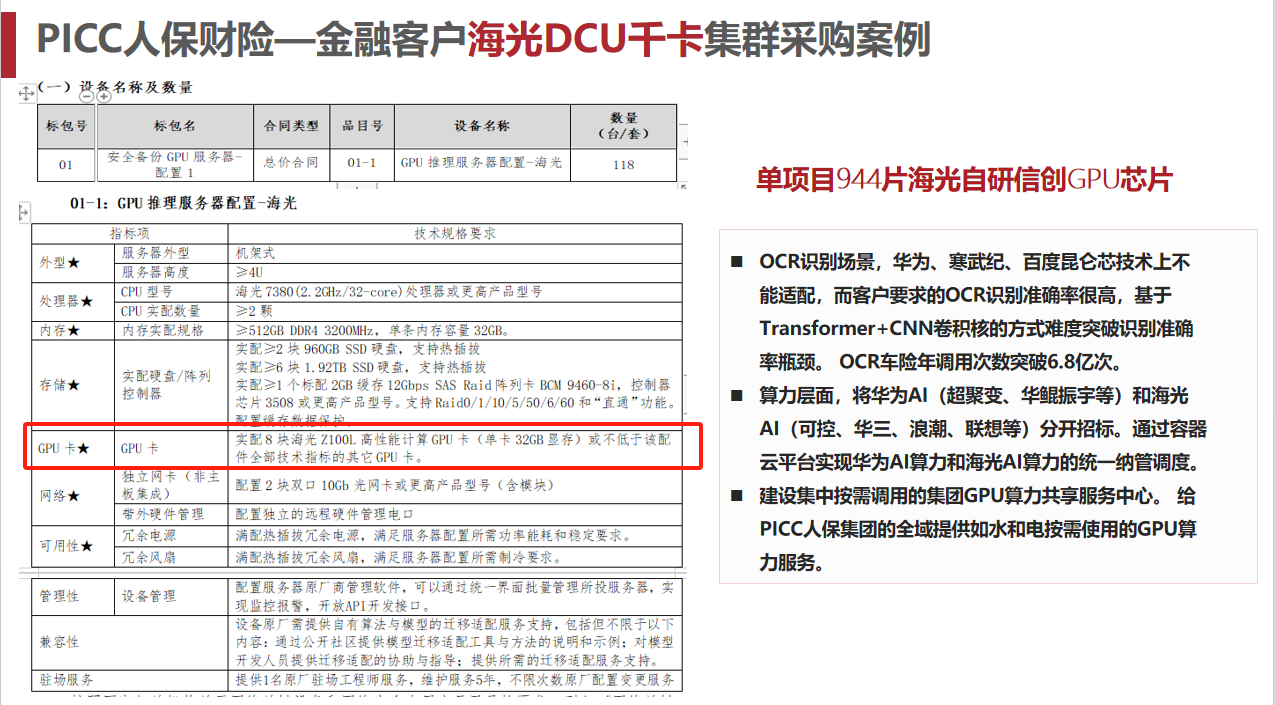
【营销系统】营销智能外呼系统，算法模型：Whisper、SpeakerDiarization等，不低于T4。

2、竞争力：

确立OCR识别优势场景。联合AI伙伴打造技术壁垒。OCR场景华为、寒武纪、百度昆仑芯技术上难以适配。 抓住刚性场景算力需求。 OCR车险理赔业务年调用次数突破13亿次，需要1200片V100算力支撑。OCR识别准确率要求高，百度、第四范式基于Transformer+CNN卷积核方式的OCR算法难以突破识别准确率瓶颈。华为、寒武纪等难以支持bert、LSTM+RNN的OCR算法引擎。

收割全部场景。组织研发力量，适配大模型、机器学习、大数据分析、RPA、客服等全场景。

一期944片：



**1.2 工商银行**

海光DCUK100-AI 满足工行关于国产卡框架入围招标需求，于2024年11月份供货212片海光DCUK100-AI 卡，主要用于OCR（文通、合合、汉王、自研）、NLP和遥感（百度）场景。

OCR场景需要国产GPU卡 支持FP32单精度推理和对LSTM模型的支持，华为非910B卡没有单精度 且缺失递归运算RNN能力，对于LSTM模型支持能力较弱，技术上未能够适配OCR场景。 2024年上半年华为正式结束对于OCR场景的适配任务，工行考虑到多样化引入和自研OCR能力，华为的生态适配过弱，决策由海光DCU接替支持。

OCR是大模型推理和场景化必不可少的能力，在RAG搜索增强生成的篇章解析、要素抽取、Enbedding向量化、AI搜索等环节，OCR成为了关键的组件。 而华为在OCR上不能够适配的短板明显，海光DCU补足了这方面的算力支撑。 如果摒弃LSTM模型，强行用transformer模型，基于华为910B 来做OCR，会导致准确率下降和算力成本大幅度增加，对于卡证、票据、文字、表格等高频OCR场景，经济效益较差。

**1.3 中国银行**

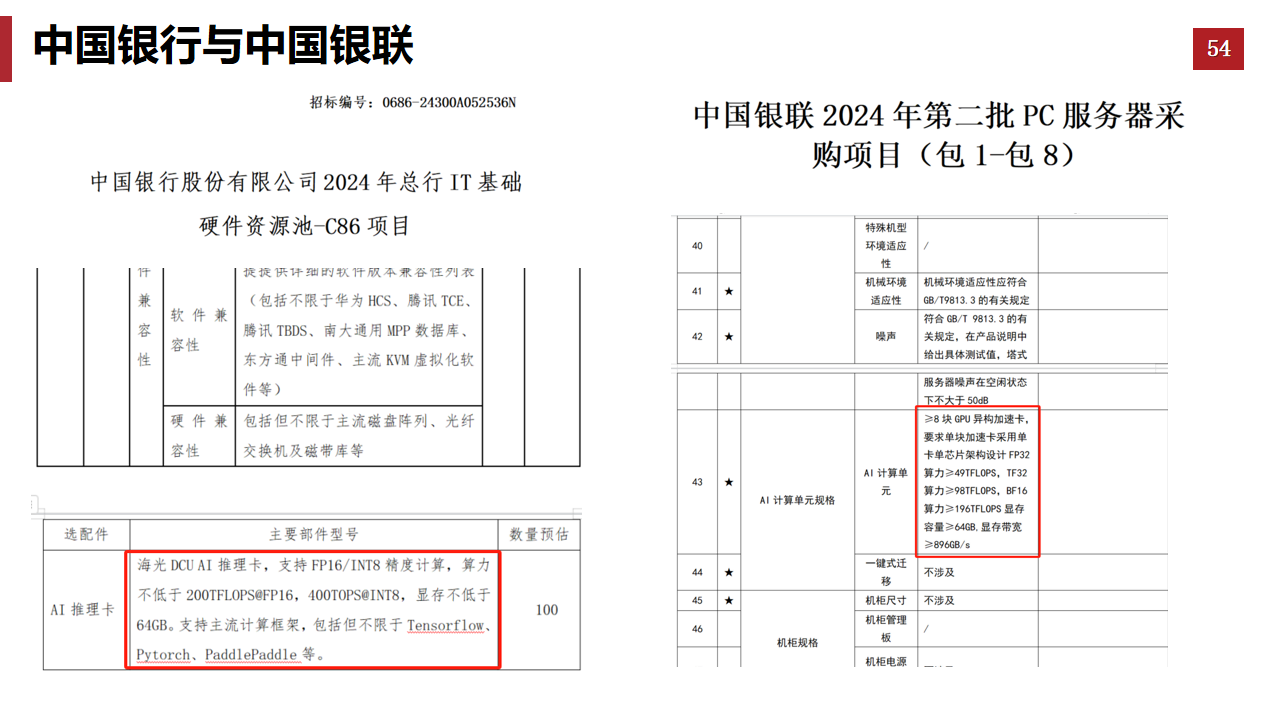
海光DCU 在2024年9月份通过浪潮服务器整机中标中国银行国产GPU采购项目，以指定品牌和型号的方式作为唯一配件写入招标文件。 主要用于OCR（易道博识、来也科技、阿博茨）、大模型（智谱、阿里通义千问）、智能客服（科大讯飞、捷通华声、腾讯）。

中行华为投标的是非910B显卡，主要用于安防监控，网点/园区的边缘侧场景。

中行OCR 三个厂家与华为昇腾的适配皆已失败，由海光DCU来支撑OCR场景算力的需求。

中行客服场景华为适配失败。客服ASR对CPU的算力需求比较高，鲲鹏难以满足，现在的方案是海光4代 7490CPU+海光DCU K100-AI的方案支撑。

中行的大模型主要是智谱中标，9B模型和32B模型，推理使用英伟达V100卡即可。 中行大模型以场景化为主，没有对训练算力的需求。 海光DCU与阿里通义千问和智谱已完成适配，且满足阿里与智谱公司商业化落地的算力需求指标。



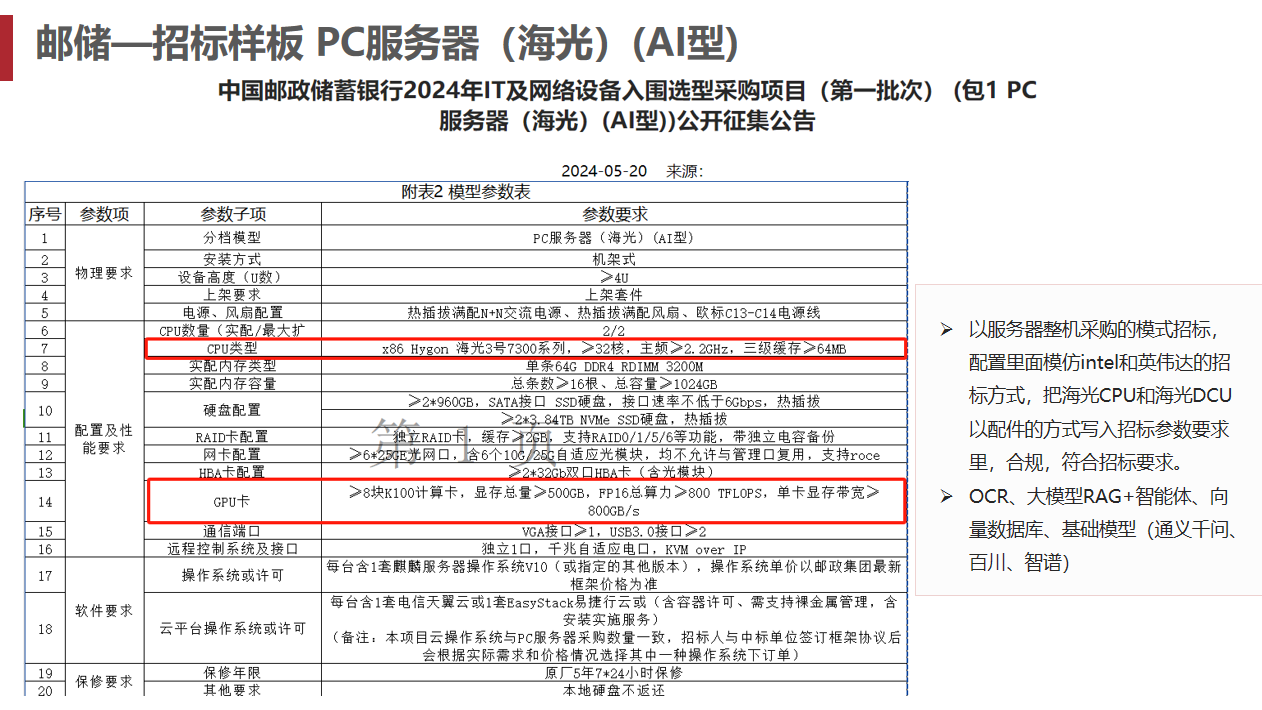
**1.4 邮储银行**

海光DCU 在2024年8月份中标邮储总行国产GPU采购项目，以指定品牌和型号的方式作为唯一配件写入招标文件。主要用于OCR（汉王、百度）、大模型（智谱、百度、通义千问）、数字人（百度）、客服（科大讯飞、捷通华声）、人脸攻防（瑞莱科技）。

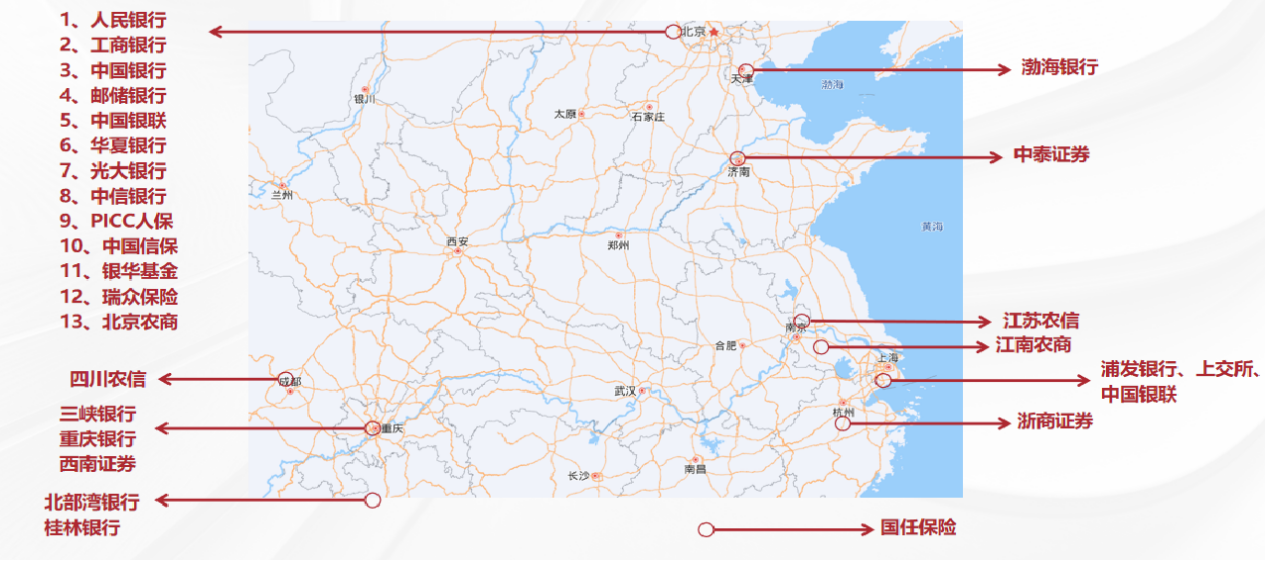
华为在邮储中标了1000多张昇腾910B卡，主要做大模型训练为主。 传统AI场景，华为昇腾难以适配，由海光DCU来承接传统AI场景的适配任务。

大模型通义千问场景，海光DCU K100-AI在72B模型可以做到4张卡量化int4推理，到22tokens/s的生成速度，超过英伟达V100卡的表现（推理速度为21tokens/s)。 满足了行里大模型推理和阿里通义千问原厂商业化的应用需求。

OCR识别场景如工行和中行，华为昇腾技术上无法适配，只能由海光DCU来支撑算力需求。 客服场景科大讯飞、捷通华声 涉及到ASR的场景，也是海光DCU在支撑算力需求。



**1.5 更多案例**



**二、海光DCU技术优势AI场景**

**2.1 OCR识别国产GPU首选卡**

海光DCU已适配合合信息、睿真科技、易道博识、第四范式、数据项素、腾讯优图、云从科技、百度、北京文通、译图智讯、致宇科技等OCR识别公司，是国产GPU中对OCR场景支持最好的异构计算卡。

技术优势如下：

1、OCR算力需求定位。

A10、V100、L20、T4 级别的卡就足够。 高算力高价格的卡对于OCR场景性价比低。

2、FP32单精度推理为主。

OCR识别需要使用FP32保障其准确率，量化到FP16及更低精度无法满足用户需求。 因此合合、第四范式、易道等主流OCR厂家均以FP32推理为主。

3、华为中低端卡无FP32。

华为只有910B（1-4）有单精度FP32，8卡服务器价格在百万级别，成本方面不适用于OCR场景。

4、CV视觉与NLP自然语言处理 双路线需要。

华为、寒武纪、百度昆仑芯在LSTM+RNN等循环神经网络的支持较弱，技术上难以优化。 DCU的通用架构对于卷积运算和递归运算支持度更好。

5、DCU已完成全国OCR主流公司的适配和合作。

联合ISV（合合信息、第四范式、易道博识）打造技术壁垒。

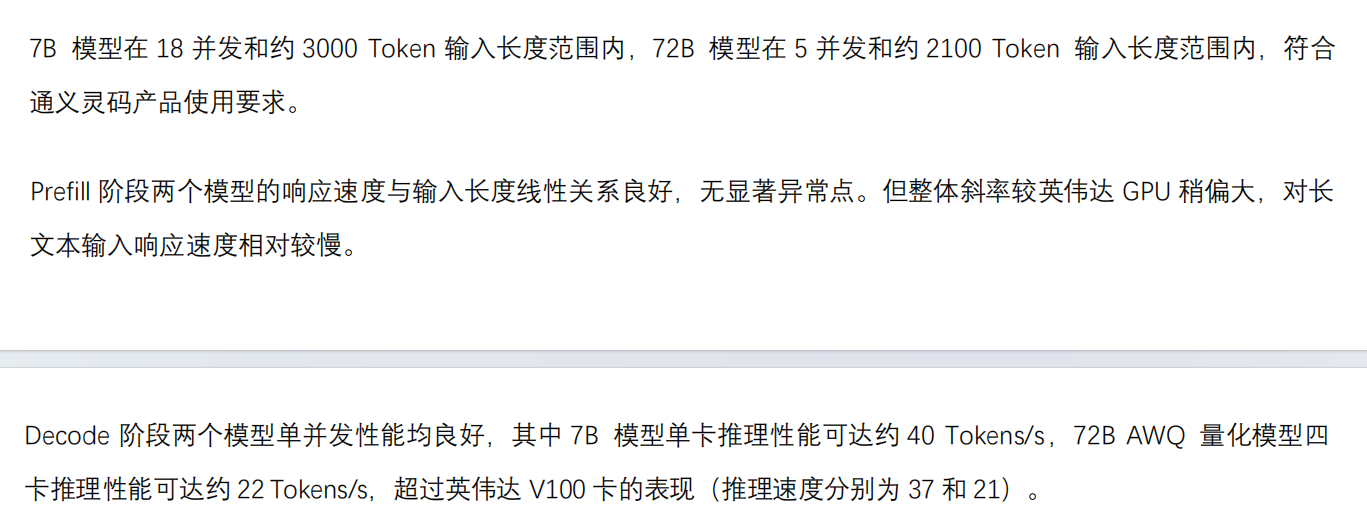
6、DCU在OCR场景已落地千卡订单。

PICC人保财险项目，DCU在OCR场景优势明显，竞对无法通过测试，唯一配件指定写入招标文件。

**2.2 大模型推理性价比最高**

大模型已适配智谱、阿里、百川、百度等大模型厂家。 与阿里通义千问、智谱公司战略合作。以阿里灵码为例，性能全面超越V100产品。

阿里通义千问72B 2.5版本。 4卡K100-AI，VLLM 0.6.2推理，AWQ量化int4，4000tokes输入，一秒返回首包，生成速度22tokens/s 。



已落地的大模型场景代码编程（阿里灵码、百度comate、智谱codegeex）、全行问答知识助手（数据项素毕昇、庖丁、阿里、融汇金信、中科金财）、大模型客服提升（捷通华声、科大讯飞、腾讯）、数字人大模型（百度、讯飞）、大模型数据分析BI、AI搜索、多模态等。

2.3 更多AI场景

1、生物识别（人脸、指纹、虹膜、指静脉）。

2、语音识别、声纹识别、语音克隆、语音质检、语音合成。

3、机器学习，大数据分析，客户画像，精准营销等。

总结：海光DCU主要用于大型金融机构的传统AI场景（生物识别、OCR、客服）和大模型推理场景，替代英伟达V100的算力需求。在OCR文字识别、ASR语音识别领域补充了华为昇腾的生态适配性差的短板。 华为昇腾+海光DCU可以完成金融数据中心对英伟达A100、V100 卡的国产替代任务。 让灵活开放的算力土壤，满足每一位开发者的AI场景化打造的需要。

1. **海光DCU竞对优势分析**

**3.1 金融客户国产卡中标情况**

2024年金融行业6+12家银行总行、4大保险、头部证券。在数据中心国产GPU市场中标情况如下：

华为中标14家。工行、交行、建行、邮储、招商、中信、民生、光大、太平洋保险、广发、平安银行、平安保险、华泰、中信证券。

海光中标12家。 工行、中行、邮储、中国银联、光大、华夏、中信、渤海、PICC人保、浦发银行、上交所、国泰君安。

天数智芯1家。 太平金科。

寒武纪1家。 交通银行。

百度昆仑芯1家。 招商银行。

沐曦电子 1家。招商银行。

摩尔线程 0家。

从中标案例分析，华为昇腾910B系列和海光K100-AI卡具有强优势。华为昇腾910B系列目标是替代A100卡，以大模型训练场景为主。海光DCU K100-AI 目标是替代V100，A10，L20，T4卡，以生物识别、OCR识别、智能客服、大模型等 的推理为主。 海光DCU性价比和生态适配性有较强优势。

其他卡商金融行业处于起步阶段，努力打造首单示范工程。 在生态适配性、大客户服务支撑团队、产品成熟度等需要一定的时间孵化。

**3.2 主要关注的技术指标：**

1、显存容量。一般来说大模型推理64G显存是标配。可以部署更多的实例，可以部署更大参数量的模型，一般来说72B量化int4 要最低128G显存。 华为64G的910B3 价格在120万以上，价格高昂。 海光K100-AI卡为64G显存，比V100、L20等的显存要大，比华为价格低很多，性价比高。

2、显存带宽。 在大模型领域比较关注，涉及到首包的返回时间。 海光DCU K 100-AI 已做到智谱、百川、阿里通义千问对首包的需求，4000tokens输入，1秒以内返回首包的需要，并与原厂已有标准镜像。 K100-AI是GDDR6，不会受到美国HBM制裁的影响。 华为的推理卡是910B3，也就是atlas 300T pro 被认为是代工的残次卡，其显存和显存带宽比910B3 要低很多，如下图。



3、FP16的指标。 K100-AI的标称值是FP16=196T，海光公司在K100-AI研发出了第一代Tensorcore计算单元，49T\*4=196T标称值。 第一代Tensorcore在持续优化中，实测的是130T左右，V100 是125T，寒武纪X8 是98T。 华为910B标称值是FP16=313T，实测在 220T左右，A100 70%的性能。

4、FP32的指标。在FP32=24T （OCR推理）DCUK100-AI 保持领先优势，是V100 FP32=15.3T的 1.5倍性能。 OCR识别是海光DCU的绝对优势场景，华为、昆仑芯、寒武纪等一众NPU芯片都做不了OCR识别场景。 主要是对于高频的卡证、票据等识别精度不足，上不了实际生产环境。

5、大模型推理量化能力

智谱、阿里、百度、字节等大模型主流厂家原厂使用的推理精度一般是int4量化为主， 如海光DCUK100-AI 在Qwen 72B模型4卡推理 int4 AWQ 量化 VLLM 0.6.2 可以到22 tokens/s的生成速度，超越英伟达V100产品。

而华为卡不支持int4精度， 性能指标 FP16 到 int4 是 1.8倍的性能提升。 72B模型， 64G 2卡（共128G）就可以推起来，大幅度节省算力。 FP16到 int8 是1.3倍的性能提升，72B 评估得 4张卡推理。 这是为啥华为910B3 主打训练场景的原因，做推理就不值150万一台8卡机器了。 华为在推理现在是主打910B4，价格降低到60万以下。

1. 卡间互联能力。 DCUK100-AI 没有卡间互联能力，在传统AI场景和大模型推理场景不受影响。 在大模型高参数量训练方向的效率比华为昇腾910B有差距。 但从光大、华夏、渤海、PICC人保等海光DCU 整包的大模型软硬件方案中，基本上已经没有了预训练的需求，微调的需求也不明显。 大模型RAG架构主要是以推理为主。 英伟达 V100和海光K100-AI 推理就够 。

**3.3 生态适配场景化支撑能力**

金融行业的软硬件分离招标，必须原厂投标是国产GPU卡厂家在金融行业聚焦到华为和海光两家的主要原因。 国产GPU卡商无法控制对于AI算法厂家的场景边界，在不同客户随机遇到不同的AI厂家和各类自研AI算法。 比较难以形成复制化效应的商业模式，难以支撑融资。 而华为和海光不靠融资。

华为和海光有信创CPU的大盘子。 华为整机伙伴超聚变、华鲲振宇、长江计算等，海光整机伙伴 浪潮、华三、中兴、联想等已在金融行业规模化落地。 华为昇腾和海光DCU可以在CPU的产业下发展，投入和风险较小。

在海光AI路线下，目前只有海光DCU 在浪潮、华三、中兴、联想 四家大的整机厂家都有了自己的机型，能够支撑金融大客户的公开招标。

在生态适配性方面。 海光DCU补充了华为在传统AI不适配的问题，支持生物识别、OCR、智能客服、机器学习、大数据分析、知识图谱、AI搜索、大模型推理等。