2025-01-09 CSIG经营月会-GPU产品进展

`powellli, 2025-01-06`

0. 背景和目标

CSIG经营月会上汇报公有云GPU产品进展，时长30分钟。

本场汇报时间为2025年1月9日周四10:10-10:40

目标：

1. 汇报GPU的经营损益和产品建设，以及对未来的期待
2. 讲清楚目标客户群体的场景和区别
3. 阐明当前的几个重大挑战及应对措施

1. 执行摘要

Page1: 标题和自我介绍

各位领导上午好，我是云产品一部李力，今天由我来汇报腾讯云上GPU产品的进展，主要包括产品介绍、经营回顾和我们看到的行业趋势的变化及相应的产品规划。时长大约30分钟。

去年1月10日，我汇报了公有云上云计算产品的总体情况，当时也提到了GPU是其中最受关注的产品，今天这个是把GPU产品展开来的情况。

Page2: 执行摘要

首先是这几年的经营损益情况。自从2023年生成式AI爆火之后，GPU的需求也随之暴涨，整个2023年都是供不应求的状况，也带来了不错的经营结果。到了2024年之后，我们明显感觉到了大模型客户开始显现出头部聚集效应，整体需求趋于收缩，同时友商的竞争也严重加剧，这给我们带来很大的经营压力。

在经营策略上，我们主要有两个措施来增加稳定性和可持续性，一是推动长约，二是寻求更多有实际业务支撑的场景。

在产品能力建设上，GPU产品主要依托于公司内成熟的技术和产品基础，比如底层的整套从机房、网络到服务器硬件的技术支撑，实现集群稳定性的业内领先，最终获得AI Infra第一名。

未来三年，由于头部聚集效应的进一步增强，大模型的需求可能还会进一步收缩，但是搜广推、自动驾驶等场景的需求仍然会稳步向前。

> 小红书、Minimax、作业帮等客户都因为包销而优先退掉了友商的GPU

> 作业帮拍搜业务对GPU是强依赖，国内从18年开始winback百度，上量黑石1.0的2080ti，后切至30和40系列消费卡。
> 在阿里和百度重投的基础下，份额从0%增至超50%，在几家云厂商中份额最高。
>
> GPU带动CPU/存储/网络/数据湖/安全等增长，作业帮 云整体收入是2.5亿+，GPU收入1.26亿（占比49%），带动其他产品1.3亿+增长；

GPU 24年收入从 1月 1.7亿 增至12月 2亿；计算整体是7亿左右。

已到货1.5万卡，已售卖1.2万（7.5K 外部客户，4.5K 自研等），预留待购买1.7K卡，剩余1.2K卡推动客户购买/匹配自研春保。
近期持续增长的主要来自：小红书/拼多多/博世/酷睿程等；对应的就是搜广推和自动驾驶场景。

序号	客户名称	收入	平均折扣	毛利	毛利率
1	小红书	482,603,517	27.5%	146,747,451	30.41%
2	美团	254,062,278	23.9%	44,035,165	17.33%
3	荣耀	144,009,195	21.0%	835,313	0.58%
4	元象	130,525,229	20.6%	13,067,098	10.01%
5	作业帮	120,189,990	27.0%	37,993,785	31.61%
6	名之梦	118,467,656	21.5%	25,434,842	21.47%
7	百川智能	116,058,252	22.8%	17,821,987	15.36%
8	智谱	92,990,765	21.1%	12,995,023	13.97%
9	拼多多	66,515,362	23.4%	17,116,999	25.73%
10	滴滴	58,115,394	22.1%	11,421,410	19.65%
11	蔚来	44,395,139	23.1%	8,059,540	18.15%
12	同花顺	40,313,317	27.1%	12,310,497	30.54%
13	贝壳	36,576,184	30.7%	17,196,856	47.02%
14	猿辅导	28,993,345	38.6%	13,834,877	47.72%
15	大疆	24,789,123	15.2%	4,486,922	18.10%
16	月之暗面	24,698,129	29.4%	2,656,267	10.75%
17	博世	24,639,071	20.3%	8,671,739	35.20%
18	TME	23,657,430	35.7%	11,123,481	47.02%
19	微众	21,053,074	61.8%	15,209,337	72.24%
20	唯品会	14,217,319	41.1%	6,430,075	45.23%

GPU 2024-2025年售卖卡的组成			
外部售卖规模（万卡）	24年	25年E	25年净增
训练（万卡）	3.37	5.78	2.41
– H20	0.75	3.0	2.25
– 910C		0.06	0.06
– A800/H800	1.47	1.57	0.1 （元象退回，拼多多计划1月回购）
– V100等	1.15	1.15	
推理（万卡）	2.0	2.9	0.9
– L20/L40/L40S	0.87	1.67	0.8
– 紫霄v2		0.1	0.1
– A10/T4等	1.1	1.1	
消费（万卡）	1.28	1.82	0.54
– 4090/4090D	0.5	1.0	0.54
– 30系列/20系列等	0.82	0.82	
汇总（万卡）	6.7	10.5	3.9

> 包销规模：2.8万卡
> 包销客户数目：包销客户数从23年的29家，提升至50家+
> 包销卡型：A800 8K + H800 6K + L20/L40 4.3K + H20 2.3K +4090 2.2K + V100/3070/A10/T4等
> 4年包销：4.3K卡，占比15%，作业帮，拼多多、滴滴、蔚来、滴滴等
> 3年包销：5.7K卡，占比20%，荣耀、智谱、大疆卓驭、名之梦、百川、智慧芽等
> 2年包销：1.2万卡，占比43%，小红书、美团、医渡云、元象、京东等；
> 1年包销：6K卡，占比22%，米哈游、地平线、中科、新浪、粉笔、唯品会、金蝶等；

2. 产品定位及介绍

Page4: 产品定位

计算产品大部分面向的是普适的业务运行环境托管的场景，在上一次汇报中，我们在产品规划上有几个场景化的方向，比如分布式云/边缘计算是把标准化的产品能力部署到用户的机房中，比如轻量云是让开发者和中小企业甚至是帕鲁游戏玩家获得类似于一个“宠物”的工具，而GPU的高性能场景则是真正体现“计算”的场景，它在大规模并行计算中有显著的性能优势。与其它几个场景最大的差异在于，GPU不仅特别贵，而且还有很高的使用门槛，并且它对性能的稳定性和持续性要求会更高。

正如刚才所说，我们在GPU的产品定位上是把它和云上已有能力和公司内部的底层能力结合在一起，即基于同源同构的原则，把云上的计算、存储、网络能力，和星星海的自研硬件、网平的星脉网络等，打造一个产品能力丰富并且稳定可靠的AI原生云。

AI原生云在现在看来是非常合理的选型，但在过去其实是有一定争议的，比如原来外界普遍认为AI infra的架构应该是要颠覆云的Infra架构，所以各种智算中心，甚至包括阿里云等云厂商，都在云的体系之外独立发展AI Infra，但其实这里不仅有大量的重复劳动，而且会导致这部分业务和云上业务难以互通。随着推理业务的发展，这里的互通就更加重要了。

GPU产品本身也是有一个较好的带动作用，短期会有配套的存储、网络、数据库等需求。

> 作业帮拍搜业务对GPU是强依赖，国内从18年开始winback百度，上量黑石1.0的2080ti，后切至30和40系列消费卡。
> 在阿里和百度重投的基础下，份额从0%增至超50%，在几家云厂商中份额最高。
>
> GPU带动CPU/存储/网络/数据湖/安全等增长，作业帮 云整体收入是2.5亿+，GPU收入1.26亿（占比49%），带动其他产品1.3亿+增长；

Page5: 市场竞争情况

市场上来看，GPU这里的竞争主要是面临阿里云和字节火山引擎的竞争。我们得益于23年的包销和多场景提升了毛利水平，并且一直延续到现在，在相对数据上还有不错的进展。

友商更加激进，不仅亏本，还面临较大的合规挑战。

Page6: 几个挑战

政策、供应、运营、竞争

运营：运营效率直接体现经营结果，跟计算别的产品不大一样

竞争：从供不应求很快变成了供大于求

Page7: 效率提升

体现云原生的价值

Page8: 竞争案例

> 【大疆卓驭】
> 背景：卓驭是目前国内头部自动驾驶解决方案商，与大众、奇瑞等多家车企合作，从L1~L3级自驾均有涉及。24～25年规划约2k卡训练集群。
> 场景： 自动驾驶训练
> 突破点：
> 1）性能提升：通过星脉网络调优、H20训练性能优化，集群千卡扩展比达96%，性能对比友商A800方案提升10%；
> 2）车图云方案：国家对自驾数据要求逐步明确，端到端训练涉及数据安全、高精地图自定义、算法训练三大环节。腾讯云行业首推车图云一体化方案，包括腾讯地图、自驾专区、训练集群，满足算法训练全链条均在合规环境下部署；

【荣耀】
背景：客户以AI手机作为重点发展方向，对手机AI包括图像识别、大模型问答、文档解析等场景。
场景： 手机AI推理
突破点：
1）推理成本优化：手机AI场景多样，视觉模型、语言模型在数据预处理阶段对搭配的CPU和内存要求不一（视觉需要多高配、语言模型需要低配）。H20/L20可灵活按需搭配CPU内存配比，匹配客户不同场景需求，平均单卡推理成本最高可下降20%；
2）效率提升：通过TACO加速，在满足客户 500ms P80 的要求下，TACO-LLM 相比 vLLM batch size并发量提升最高 7.5 倍，吞吐最高提升 110.14/14.74=7.4 倍

【滴滴】

背景：客户大模型业务主要用于内部业务效率提升，期望使用H20作为训推一体集群。

场景： 大模型推理

突破点：

1）推理加速： H20+TACO大模型推理加速，在qwen-70B的在线推理测试中，实测H20吞吐提升7倍。

招商银行

背景：招行TCE项目采用GPU池化管理方案，通过qGPU进行了资源细粒度切分，可实现从0.1卡到千卡的灵活调度。24年国家提出国产化切换的新要求，招行需使用多厂商国产化芯片。
场景：人脸识别、智能客服等业务
突破点：
1)利用率提升：通过qGPU统一调度管理资源，GPU资源利用率提升30。qGPU作为重要的提高客户粘性的手段，基于TCE平台广泛用于客户T4、A10、V100等卡型。

> 2)多芯兼容：qGPU进一步支持NV、沐曦、昆仑芯等多平台池化调用，客户可在不改变原有调用模式的前提下使用国产卡，作为重要差异化竞争点；

> **Liblib**
>
> 场景：文生图推理平台
> 突破点：
> 1）TKE原生节点集成：通过原生节点，客户可实现CPU/GPU资源灵活混部，并利用上腾讯自研“悟净”内存管理技术。显著提升GPU资源利用率，降低客户人力运维成本；

2. 经营回顾

Page9: 经营回顾

做好了库存的快速销售，并且确保了行业的多样性和经营的健康性。

头部客户定义：GPU年收入的1%，即GPU年收入超过2KW。19家头部客户占比80%收入

分类	增量（H20等）	存量	合计
外部售卖规模/万卡	3.06	7.28	10.33
25年收入/亿元	8.73	21.85	30.58
25年毛利/亿元	0.46 [5年折旧 1.70]	6.39 [5年折旧 7.05]	6.85 [5年折旧 8.75]
25年毛利率	5.3% [5年折旧 19.5%]	29.2% [5年折旧32.3%]	22.4% [5年折旧28.6%]

3. 未来规划

Page10: 重点场景

大模型有了头部聚集效应，这两年的主要需求来自通用大模型的训练，但这部分是在收缩的，未来会在垂直大模型和大模型推理上有一些进展。这个领域的特点是市场竞争最激烈，很多客户也不大在乎合规的要求。我们的策略是保住存量规模，尤其是2023年最热的时候那些通过包销锁定了客户的规模，未来更多提供一些竞价和短套餐方案来对冲库存。

自动驾驶场景是一个非常高速增长的场景，并且相比大模型，它更聚集真实业务场景，且客户付费意愿更强，而且更在乎合规性，是我们主推的新场景。我们通过自动驾驶专区，推进H20的售卖。对别的产品有较好的带动作用

搜索广告推荐是最有实际业务支撑的场景，它能直接给客户创造经营收益，而且客户也基本上是云上已有的那些超大客户，比如拼多多和小红书。客户基本上是把业务系统跟着GPU部署的，给他们提供GPU对于防止客户流失意义重大。

传统AI比如OCR/ASR等也是稳定发展的场景，它们对于推理卡有更好的接受度。比如作业帮和猿辅导等客户也有很大的规模增量。

从右边这个图里可以看出，大模型的占比在逐渐降低，增加了对于更稳定可靠的业务场景的渗透率。

同时，也可以看到GPU直接带动的产品收入的毛利是更高的，因为GPU占比大，客户更容易直接比价，但别的产品我们有更好的议价空间，比如存储我们的云原生理念也确实有显著的优势，也有了更健康的经营情况。

Page11: 产品规划

基于保持对云原生和发挥腾讯自研能力的坚持，我们的产品规划是做好与自研的资源协同的技术协同，尽可能提升我们的整体资源利用效率，也帮助用户提升他的资源利用效率。

大体来看，自下而上有这么几个规划：

1. 从部署方式上，考虑到一些用户就近接入的需求，还有一些用户自己有卡而运营不起来。我们可以把整个上层的软件套件可以以分布式云即类似于边缘计算的方式输出，也可以以专有云的方式输出。
2. 在产品能力提升上，我们计划在资源和技术上与公司的资源打通，寻找更多弹性使用的需求场景，来提升资源的利用率。同时我们也在算力服务化和硬件迭代上进一步将性能提升，客户的易用性提升。
3. 在框架上，以前GPU提供的界面有限，未来我们会将加速框架和Tione等。

Page12: 未来三年

未来三年，我们预测复合增长率是18%。我们看到了整个市场已经处在一个激烈竞争的阶段，但我们相信AI的发展仍然有很大潜力，通过多场景齐头并进，尤其是关注有实际业务支撑的场景的加入投入，提前规避经营风险，做好对GPU大盘经营的保障。同时也希望通过GPU守住了其他产品的规模，并进一步促进用户将更多业务部署到腾讯云上来。