25025 计算产品bsc-coby汇报

`powellli, 2025-06-16`

背景

云线宁夏offsite汇报年中bsc，汇报时间为6月17日上午11点以后，时长15～20分钟

Page1 经营损益

1. 计算在2025年H1的收入是42.48亿，同比增长9%，完成率99%，没有达到100%的原因主要是受到年初支援deepseek api的影响，GPU的完成率只有92%。
2. 全年收入预估可以达到88亿，100%完成目标。因为一方面上半年我们看到了一些去年做bsc时预期以外的客户落地，另一方面GPU的对外供应也恢复了。当然这里最大的风险仍然是GPU的市场不确定性，后面会展开这里，刚好也是借此机会请各位老板关注。
3. 在毛利层面，H1毛利是21亿，同比增长37%，完成率136%。这里与预测的偏差较大，其实主要原因是除了GPU以外，对损益有一个大的正向因素和一个大的负向因素，负向是市场竞争的恶劣程度和大客户的退换节奏超出预期，正向是我们有了更多的新增的大客户补齐了这里的缺口。同时因为新增的核数变多，那我们在成本上的优化就发挥了更大的经营结果。
4. 全年毛利预估可以达到39亿，全年同比增长22%，完成率124%。下半年预估表现下滑的原因是考虑到大客户降价的影响以及618之后存在的潜在退还情况。

5. 过去几年，我们受到的挑战总结起来其实就是：**我们的一两个大客户的占比过高，而它们使用了定制化的机型，并且有不可控的退还行为**，其实就是经营结构的风险太大。在大家的帮助之下，总的来说，当前时间来看，计算产品的经营结构健康度提升了很多：
    1. 客户结构上：以收入超过千分之五的客户定义为大客户的话，从原来的1家经过几年，在今年年底会有24家。**在这里我们也看到了海外的巨大空间**。
    2. 机型设计和库存管理上：依托星星海和运营的支持，我们把原来几十款机型压缩到几个。并且行业的前端同事和产品一起付出了很大努力说服客户改变原有理念，用我们的标准的高密度低成本机型。这样可以充分发挥供应链的成本和交付优势，进一步提升经营效率。同时也有利于整个大盘的利用率提升。
    3. 长约锁定上：基本上大客户都已经签订了包销，有些因为市场竞争无法确保的，也和客户开始有了默契，可以在提前较长时间通知的机制下减少轮转带来的损耗。

> 平账其实更多的是错期，真正的平账占比不到5%

Page2 损益细节

简单分开通用计算和GPU来看

1. 可以看到CPU(通用计算)这部分相对比较稳定，跟预测的数值也差不多。当然利润因为新增用户的原因是略超出预期。
2. GPU我们希望在H20到货之后能在下半年补齐上半年的缺口。毛利的异常升高是由于GPU的成本折旧从4年变成了6年，所以会有一个短时的很大的利润提升。这些账面上的利润提升也完全打到行业这里的客户级损益了。但是我觉得这里的风险仍然是非常大的，H20很有可能无法支持

6年的完整使用，我们得想办法把卡卖给出价最高的客户并尽快把成本收回来，这样到了第5年第6年万一不具备技术竞争力时就不会那么难受了。

> GPU还原到4年折旧，毛利会从163%完成率变成102%完成率，对大盘影响不大
>
> GPU还原到4年折旧，毛利会从113%完成率变成108%完成率

Page3 下半年损益提升

我们从国内和海外分开来看对市场的感知和重点的突破方向。

1. 总的来说，海外收入占了大盘11%左右，毛利占了大盘15%左右。但是新增收入有30%来自海外。
2. 从图中就可以看出，市场竞争在国内和海外是完成不一样的。我们国内一共38亿的收入，因为国内太卷，躺着不动就会掉5.75亿，这都有点冲击所谓"订阅式收入"的逼格了。我们靠大客户和GPU的收入补齐缺口做到增长。
3. 国内的市场竞争的影响是15%，海外只有6%，如果以此来看，国内的竞争激烈程度是海外的3倍左右。今年海外Garena, Bybit, 美团都已经突破了10万核的规模，我看数据应该futu今年也很有希望突破，这都给我们很大的信心。同时Goto、正大这些海外特别标杆的客户也在持续上量了。

Page4 大客户

大客户这里其实就是围绕着**一两个大客户的占比过高，而它们使用了定制化的机型，并且有不可控的退还行为** 这个问题。

1. 我们先看左边这个图表，可以看出规模其实一直在增长，但受限于市场竞争，总收入的变化其实并不大。当然另一方面，收入是规模的后置指标，订阅制的特点是规模先于收入增长（或下降）。
2. 以千分之五以上收入作为大客户标准的话，20/21年只有一两家，23年8家，24年18家，25年现在是22家，年底可以到24家。大客户变多，互相拆借和抵御风险的能力就能变得更强。
3. 在产品能力上，首先要讲的肯定是AMD机型，这个我认为是当前产品上最大的优势。过去两年跟前端的交流产品roadmap，我都是说只要让客户接受了SA5，那结果就一定不会差。到现在SA9的时代，我们已经看到整体技术和产品能力都是在业界保持非常大的优势。这里有非常多的技术突破，比如星星海的散热技术支持只有我们能提供3.4G睿频，可靠性这里有高密RAS建设，内核这里支持256核以上的虚拟机，高密机型独有的亲和性要求在装箱算法上的优势。
4. 小红书从原来不接受AMD，到不接受AMD高密，到目前已经打算全面切换SA9，是靠战略客户部带着后端的专家们不断攻坚突破的结果。在小红书需要核性能更高，单核内存带宽更高的场景中，我们通过SA9关核的方案做到不增加供应新的机型情况下，让小红书也获得更多的虚拟机机型并有足够的性价比收益，预期Q3有机会在规模上反超阿里云。目前了解到阿里云在紧急引入高密AMD机型。
5. 除了亲和性调度保证虚拟机的性能和稳定性以外，对于整个大盘的调度也在过去几年的库存优化中得到了较大的提升。一方面通过混部增加调度域，一方面在运管端到端资源管理的带领下，整体迁移能力从原来的每月20万核提升到80万核每月。以京东外卖为例，满足了京东50万核的突发需求。是首次真正让京东使用腾讯云，而且因为我们在高密上的优势，京东发现腾讯云的优势远大于京东云。
6. 调度能力的第三个方面是大客户的精准服务，然后想办法把算力的潮汐利用起来。

7. 潮汐算力在国内最早是火山引擎提出来的差异化竞争，他们声称每天可以提供上千万核的抖音退出来的算力，给到渲染相关的客户使用。我们利用大盘本身的潮汐特征，在产品上支持竞价实例，固定时段实例，过保专区实例等。当前渲染客户的弹性核数在国内是第一名。

> ● 本周腾讯云潮汐售卖核数已经冲到160w核（还有冲高的趋势）；近期均值在110-120w上下浮动
>
> ● 火山引擎规模在110w核（峰值）平时70w核上下浮动；
>
> ● 阿里云常态在100w核浮动；峰值能达到150w；

Page5 GPU

GPU是风险最大的部分，因为GPU无论是供给还是需求都有更大的不确定性：供给受限于国际环境的变化，很难预测；需求被火山和阿里卷得不行，而这两家又有非常强烈地需要靠GPU收入来吹大收入规模的目的，在出价上非常激进。火山最近甚至开始在承诺用户以竞价实例购买H20，但他们保证不会回收。

1. H20最重要的策略是要有面对不确定性的快速有效响应。因为它属于稀缺资源，所以在CPQ审批、分货上都会更严格一些。总体的设计原则是价格高的，使用时间长的，能带动更多产品收入的商机会排得更前一些。当然也会存在一些弹性的，这种就是反过来希望用户接受可回收的场景。
2. 由于我们把GPU放在云的体系当中，除了获得更好的AI Infra的各种产品能力以外，通过混合调度的方式，也可以提供不同CPU和GPU配比的虚拟机机型，对于一些激烈市场竞争的情况下，也许我们可以通过降低CPU数量来降低成本，从而获得更好的竞争优势。以美团为例就是这样。
3. AI Infra也在继续发挥它的价值，补齐客户工程能力不足。

> 1. 长安SGLang推理框架支持InternVL模型。InternVL是上海人工智能实验室的“书生-万象”
> 2. 酷睿程(大众和地平线合资公司)A100切换到H20，FP64算子切换到FP32。
> 3. TACO在DeepSeek-R1 满血版场景下，相对于荣耀原始线上业务性能，TTFT（首Token延迟）P95的响应时间最高降低6.25倍，吞吐提升2倍，端到端延迟降低 100%。在社区最新版本 sglang 场景下，TTFT P95 的响应时间最高降低 12.5倍。模型运行更平稳，系统调度更顺畅。
> 4. 大幅提升SGLang框架大模型服务冷启动的模型加载速度。

Page6 出海

1. 从收入和核数增长的对比可以看出，海外目前还处在较为良性的市场竞争环境当中。
2. 过去增长得还不错，得益于整体能力的提升。
    1. 在裁撤掉严重亏损机房之后，基础建设方面拉齐了国内的基线，以高标准建设。
    2. 资源交付层面供应链体系在VMI和机位优化上大大提升了效率。
    3. 产品能力补齐。
3. 但是仍然看到很多产品方面的不足。海外，尤其是海外local的客户，跟国内还是很不一样的。他们普遍对文档、API要求更高，而这方面我们有不少工作需要补齐。
4. Supercell原来觉得腾讯云不行，因为文档无法闭环完成。真正接入之后发现性能稳定性都非常不错，他们建议我们应该重点把文档体验补齐。

1. CapitaLand：凯德集团，亚太知名地产公司，总部在新加坡。国内部分：主体是凯德中国，2024年与腾讯云签约，将国内业务从Azure搬迁到腾讯+阿里，预计25 年年底全部搬迁完成，腾讯云占比30%，300wRMB。海外部分：海外主体，合同正在走审批流程，价格和方案已讨论完毕，切Azure30%到腾讯云，总金额400 万RMB，计算部分50%，200万-250 万左右，全部按量+SA5。

2. Amagi：全称 Amagi Corporation ，北美客户，为电视和流媒体平台提供端到端的SaaS解决方案。目前计划将部分非核心业务从AWS迁移至其他云厂商，以达到降本诉求，邀请了阿里云、腾讯云两家公司参与报价。项目金额约1700万RMB/年，计算占比预估30%。
3. eMAG：罗马尼亚最大在线零售商，预估量级1万核左右
4. 泰国Central Retail：泰国知名零售集团，第一期几十个超市门店，预计2M一年
5. 香港赛马会：1.5M一年

SA5单核收入新加坡48 曼谷 44 香港 72，北美65，法兰克福 54