2025年中REVIEW

`powellli, 2025-06-24`

基本内容

1. 自我介绍
2. 负责版块及业务产出
3. 未来规划和建议
4. 有什么问题

1. 自我介绍

1. 2011年的校招生，一直从事腾讯云IaaS产品的工作
2. 从小认为信息的远程与快速传播是一种魔法，由此7年的通信系统的高等教育经验
3. 2008年汶川地震，通过QQ帮助班上同学报平安，对腾讯有极其强烈的认同感

4. 毕业前夕，认为互联网的技术将要建设一个数字化的世界，希望来到世界上最好的互联网公司成为世界上最好的程序员
5. 加入腾讯B2以后，开始作为草创成员从0到1搭建腾讯云的底座，历经计算、存储、网络等多个技术路线的经验
6. 感谢公司给予的非常开放和务实的工作环境，得以有机会在2012年设计和实现了腾讯云核心调度系统
7. 2017年，出于义愤在外网发表过一系列为云正名的文章，结果意外被上级建议转型产品负责人。这个任务对我挑战很大，学习了如果从“对机器编程”到“对人编程”，思考如何从产品的角度看待tob市场的技术能力、服务能力的平衡。
8. 2020年开始，面临经营上的巨大挑战。某超大客户大规模退还设备，不仅给收入造成断崖式下跌，退还的库存也成为了损害利润的主要因素，毛利上从微弱赢利急转而下变成负数，而在2021年以后，几乎整个互联网行业的客户都开始降本增效，进一步加剧了经营上的问题。
9. 经过几年死磕，毛利从最差时候的-50%提升到40%以上了。今年能完成88亿收入和39亿毛利。
10. 目前经营结构得到极大改善。1. 客户占比更健康；2. 产品定价和长约锁定更完善；3. 技术突破大幅降低物理机型数量；4. 弹性能力促进资源运转效率。

> 我曾经是枕着牛顿传记入睡的“小镇做题家”，因为对“熵可度量信息”感到惊奇而学了七年通信工程专业，毕业后怀揣着“用技术改变世界”的梦想来到了腾讯云。在这十三年中，从研发到产品到经营，看到了从第一个核的运行到第一亿个核的交付，也让我从“对机器编程”的专业和务实，学习感悟到了“对人编程”的真诚和热情。云计算已经逐渐成为默认的基础设施，希望我能继续在这个浪潮中与业务一起相互成就。
>
> ——摘自2024年的《腾讯干部发展自评》

2. 负责版块和业务产出

主要负责计算产品、区块链产品、域名解析产品

1. 计算产品是云产品中最底层的部分，是大部分客户上云使用的第一个或唯一产品，收入规模约为整个云大盘的四分之一，并且带动存储、网络等产品
2. 计算除了通用计算作为坚实大盘以外，还有GPU、轻量云、分布式云几个方向，这几个新的方向在产品和技术上均是国内第一名。
    1. GPU的成本高昂且使用门槛高，对云上产品提供的能力要求更高，我们在很早开始成立专门的团队，帮助用户用好GPU。其中最重要的一个设计准则是把GPU融入到云的体系当中，充分发挥已有优势，减小投入和风险。
    2. 轻量云即以LightHouse为核心的产品体系，云计算越来越复杂，反而提高了对个人开发者和中小企业工程师的使用门槛。我们通过基础产品的功能裁剪和整合，提供开箱即用的公有云服务，也得到了社区的广泛欢迎。这个产品为腾讯云赢得了“良心云”的美誉。“帕鲁之战”体现了我们的强大产品能力。
    3. 在产业互联网这个层面上，客户上云的阻碍通常在于本地化的服务以及与原有it基础设施的整合。我们率先提出了管控面和数据面分离的模式，以更灵活和产品化的方式提供了云往边缘延伸的技术架构，这种方式在业界一般叫做“分布式云”，它非常好地解决了传统企业数字化转型过程中对于数据管控的担忧，并且以混合架构提供标准且灵活的用云体验。

3. 区块链产品是一个较小投入的补全性产品，用于产业数字化过程中的数字确权，同时承担了公司战略项目“长安链”，与微芯和浦芯合作，目前已经承担起约70%的长安链研发工作。
4. 域名解析产品支持了公司所有自研业务的解析，同时也在云上支持用户。这个技术过于古老，反而一直被忽视。由于阿里云的屡次故障，我们需要进一步夯实基础能力。

> **业务场景分类**
>
> 第一类是目前我们的最主要收入来源，即互联网的大客户群体，他们看待云服务器就像是看待工厂里的流水线生产装备，被集群式地应用在解决具体功能模块的计算需求，所以客户对产品的需求主要是高性价比、高弹性等。我们将这类需求认为是computing。稳定性和灵活性
>
> 第二类是来自于更广泛的产业互联网用户，在使用方式上跟前一类其实差别不大，但是在视角上，云服务器更像是一个业务托管平台，所以客户对于这个平台的位置、是否符合行业标准，以及由于it技术水平相对偏低对于产品间的集成和服务水平有更高的要求。我们称之为hosting。可扩展性和合规性
>
> 第三类是以个人和小团队开发者为主的群体，他们把云服务器当作“宠物”一样看待。客户的开发、测试、部署运维都是在同一台设备上完成，并且这些用户更多出于个人爱好和小型业务需求，他们会对易用性要求更高，希望产品减少技术概念，快速切入到用户自己的业务当中。易用性

> GPU
>
> AI云原生，就是以云原生的整套体系，支持AI Infra的发展。过去十几年，全世界的整个IT技术生态打造了以云为中心的基础设施最佳实践，这就是云原生的体系。我们在AI的浪潮中，坚持这一理念，在云原生的基础上提出了AI云原生的概念，也就是将云原生与AI的需求结合在一起，发挥云原生的标准化和技术生态的优势，将AI训练和推理应用更快更好地落地。
>
> 从AI云原生，场景上也从前两年的以LLM大模型训练为代表的场景延展到如今智能体、智能驾驶、机器人等更广泛更贴近实际应用的场景。

> 轻量云
>
> VPS本身是一项非常古老的产品，它最早可以脱胎于1970年代Unix大型机的共享目录，每个用户拥有一个文件系统上目录的读写权限，将web页面存放到某个路径，即可在web上访问。
>
> 从技术发展的角度上看，业界通常认为，随着云厂商的虚拟化和大规模调度技术的领先优势，VPS将逐渐式弱。
>
> 但事实上来看，业界的VPS仍然具备旺盛的生命力，以海外为例，digitocean在2012年才成立，但是快速就获得了100万以上的活跃用户，甚至超出大部分云厂商的营收。
>
> 这给我们的思考在于，云计算的发展越来越往高度工业化去发展，而VPS这种性能和稳定性偏差，但是注重web托管和个人开发的形态仍然具有巨大的竞争力，广大开发者、技术爱好者和业务形态较简单的中小企业在VPS很容易上手，并有一定黏性。
>
> 我相信，我们的友商也是看到了这个原本并不起眼的机会，AWS在与VPS厂商竞争利的情况下于2017年初率先推出Ligthtsail，并迅速扩张，阿里云在当年也快速跟进，推出阿里云轻量应用服务器，快速获得国内领先的地位。

> **分布式云**
>
> 控制面和数据面分离的模式，在技术上的挑战不大，在产品上却非常好地拓展了新的增量场景。
>
> 尤其是对于工业和能源等传统行业来说，它们倾向于在自己的机房或者业务所在地部署业务，而不擅长做基础设施的管理和维护。分布式云的方式同时满足了这两个诉求，通过复用公有云的管控能力，将计算节点延展到业务需要的地方。
>
> 这里值得一提的是，Gartner今年也把行业云作为未来云的发展方向之一，行业云就是针对特定行业的业务特征形成的具有行业属性的云平台。我前段时间去上海跟Gartner的分析师kevinji研讨的时候，他们提出一个新的观点，我们在宝武的分布式云的合作，意味着通过分布式云来承载行业云会是一个很值得探索的方向。
>
> 核心设计原则其实就两个：
>
> 1. 本地部署
> 2. 复用公有云的能力
>
> 公有云的上层产品能力，只需要做一个特殊zone的适配，工作量也很小。

几个业务产出可供展开：

2.1 通用计算通过强大的调度能力保持行业领先

大盘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。

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。

调度能力三大优势：

1. 高密机型的亲和性和反亲和性调度：

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。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。

2. 大盘的混合调度：

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。

3. 潮汐调度

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。

2.2 坚持AI原生云理念，提高GPU产品的竞争力和市场优势

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。

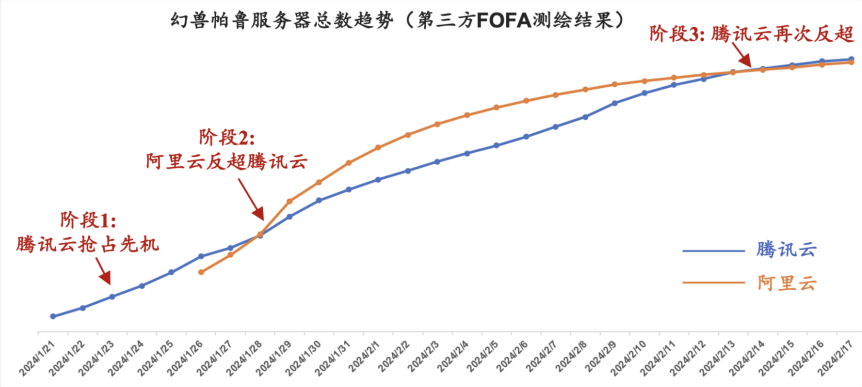
具体卡型损益预估

卡型分类	卡型	当前计费规模	Q1 收入	Q1预估毛利	Q1预估损益	预估年底规模	全年预估收入	全年预估毛利	全年预估损益	
训练	H20	14,000	100,030,254	12,303,721	12%	30,000	940,193,420	122,225,145	13%	
	A800	7,974	126,998,192	32,383,415	25%	7,974	431,793,852	99,093,249	23%	
	H800	6,862	210,241,057	66,223,453	31%	6,862	714,819,594	202,643,767	28%	
	V100等	11,636	54,474,132	12,025,319	22%	9,636	185,212,050	102,257,653	55%	陆续过保
	小计	40,472	491,743,635	122,935,909	25%	54,472	2,272,018,915	526,219,814	23%	
推理	L20	10,602	39,704,260	7,134,359	18%	33,300	294,682,390	58,936,478	20%	
	L40/L40S	2,872	33,856,107	7,133,965	21%	2,872	115,110,764	21,829,934	19%	
	A10	2,626	11,597,053	2,631,639	23%	2,626	39,429,981	8,052,815	20%	
	T4等	10,526	34,743,886	6,263,156	18%	8,628	118,129,211	40,745,700	34%	逐步过保
	小计	26,625	119,901,306	23,163,119	19%	47,425	567,352,346	129,564,927	23%	
消费	4090d	4,066	10,807,525	1,513,053	14%	10,000	79,088,050	7,908,805	10%	
	4090	3,110	14,640,992	4,450,862	30%	3,110	49,779,373	13,619,636	27%	
	30系列/20系列	7,499	7,165,362	-908,764	-13%	6,565	24,362,230	2,466,904	-11%	逐步过保
	小计	14,675	32,613,878	5,055,151	16%	19,675	153,229,653	23,995,346	16%	
总计		81,773	644,258,819	151,154,179	23%	121,573	2,992,600,914	679,780,087	23%	

2.3 发挥云的规模效应，打造开发者喜爱的“良心云”

以帕鲁为例：

> 从一场节奏紧张的演习变成一场残酷撕杀的战役，几乎每个小时都有新的情况，似乎有点像是俄乌战争的剧情，现在来看我们以相对可控的投入获得了大量非常宝贵的经验和教训，以及更重要的是也让我们获知了对手或令人敬佩或令人惊讶的竞争策略和经营数据。
>
> 整个战役大概分以下几个阶段：
>
> 1.21-1.23，调研和启动。制定了以应用创建、应用分发、应用管理三个方面的产品能力框架，目标是最终售卖10000核。
>
> 1.24-1.28，产品功能快速适配。幻兽帕鲁游戏在线人数持续攀升，阿里云在不到60小时后反应过来并直接发起价格战。产品功能集中在一键创建、便捷配置和安全防护几个优势上，售卖规模超出目标十几倍。但是也因为阿里云极低价格和不计成本的全网推广使我们丧失了规模上的领先地位。
>
> 1.29-2.3，产品运营全方位竞争。保持产品优势的同时，多团队联合增加全方位的运营推广，十秒开服、可视化配置面板、游戏存档备份和迁移、高阶安全功能成为用户认可的产品调性，在运营稳定性上保持性能稳定且没有故障。游戏热度开始降低的情况下保持规模继续翻番，减少与对手低价策略的总规模差距。
>
> 2.4-至今，接近尾声的新开始。游戏热度加速下滑，产品上继续以3秒开服、混元AI助手、存档云备份等方面探索更多极致地体验，运营上邀请游戏主播直播开服、增加海外的运营和合作等更多探索和尝试，为未来更多的轻量应用场景铺平道路。在能力保持优势的情况下与对手的规模差距进一步缩小，从趋势上看有希望能在春节期间再次反超。



	腾讯云	阿里云
产品形态	轻量云体系	计算巢
产品团队	IaaS、安全、官网、市场等团队及宝塔	计算巢主导，官网、市场等团队协助
产品体验	一致	割裂
工作量	2.7人月	预估超过20人月
营销费用	少量	预估100倍以上
产品创新	AI助手、TAT	应用面板
经营结果（首月）	收入：1200万 毛利率：20%	预估收入：1000万 预估毛利率：0%

2.4 分布式云作为高度标准化的边缘延展，拓展更多产业互联网用户场景

首先是技术架构坚持同源同构，这是减少工作量和使用风险的关键原则。

其次是对客户的界面标准化，我们还是以公有云的标准api服务本地计算的用户。

第三是在经营层面强调可持续性，用于减少经营风险，提升重资产的运营能力。

最后是在服务的过程中增强产品的优势能力，让我们的分布式云有更好的市场竞争力。

2.5 长安链已经事实上成为“国家链”，为数据要素打下坚实基础

区块链预计去年完成大约5000万的收入，毛利在4300万左右，分别较去年增长14%和23%，增长低于预期的主要原因是政府相关的区块链项目由于经费的原因减少了很多，仅从公开招标的情况来看2024年减少了50%以上。

目前我们有43个子公司同事在全职支持长安链发展，主要聚集是在区块链底层的核心技术上。从工信部的数据来看，国内的联盟链已经是长安链占绝对主导地位，并且也得到了较高的官方认可。但由于长安链本身开源开放的承诺，绝大部分存量区块链的替换是客户自己操作完成的，没有很好给我们和微芯带来商业化收益。

集团for商业化的员工hc，因为暂时缺乏明确的业务增长预期，已经由47个hc减少到了16个，未来计划会以不超过15个hc支持一个最小化投入。

从更大的层面来看，长安链的投入带来了北京和上海的一些腾讯云其它产品的商机落地，尤其是最近上海徐汇区的区块链带动的整体政务项目，后续预计在北京和上海仍然会有一些这样的机会。我最近跟董进院长沟通也跟他达成一个初步共识，即我们在长安链的大项目上通过腾讯云的多个产品和微芯的区块链产品结合售卖，应该对双方都会有更大的帮助。

政务类项目：

1. 微芯给我们带来的是大的政府项目。北京区块链云和上海区块链云，这些是微芯的区块链加上我们的TCE等软件。这些场景偏向于政策引导性的大型政务项目。
2. 基于长安链在政策方面的影响力，腾讯云自拓的政务项目，典型的是徐汇区的区块链项目。

央企/国企项目：

1. 招商局集团的航运贸易链。支持招商局的全球航运体系的数据可信交换，同时也带来了上亿的腾讯云其它产品收入。这个项目最大的意义在于被网信办评优，并且现在多个部委推动其成为国家级的航运贸易链。
2. 华泰证券、东北证券、南京银行的区块链数字身份应用。通过长安链实现金融行业身份认证，并获得证监会业务创新试点推广。
3. 建设银行、深圳数据交易所的司法存证和数据要素。在长安链之上做司法存证和数据交易的底层支持。

过去几年，腾讯云投入50人技术团队，与微芯研究院持续建设长安链底层平台ChainMaker，2024年新发布开源版本3个，已累计发布开源版本21个，开源项目100+个，代码量300w+行，技术支撑300+重大应用。 长安链开源社区生态繁荣，已汇聚50家央国企等联盟机构应用场景建设者，20+硬件生态伙伴，100+应用开发商，50000+社区用户。

24年腾讯团队积极参与由浦芯/微芯
主导全新开源品牌 ChainWeaver 的建设，已发布开源版本 2 个，开源项目 45 个，代码量超过 100 万行，在国家级场景建设中， 已形成较丰富的开源技术产品生态体系， 包含 10 大产品线 22 个二级产品， 已支持多个落地场景。 包括由中央网信办牵头，上海大数据主导，与中远海、中国银行、人保、招商局、欧冶金服等航运贸易生态企业共同推动航运贸易场景，数字货币研究所的数字人民币链上支付场景；上海市大数据局与各方数据要素流通场景以及上海市公安局的隐私计算场景； 下一步将积极探索与腾讯各业务结合的机会。

未来，腾讯云TCE、TDSQL和TencentOS等产品，与互联网3.0操作系统ChainWeaver紧密结合，共同体现了腾讯在履行社会责任和践行科技向善理念方面的决心。
腾讯云TCE作为算力基础设施，连接了上层应用与底层硬件，为国产化替换提供了坚实的基础。助力多个行业实现数字化转型，保障了国家信息安全。
TDSQL拥有安全可信的数据库解决方案，产品技术领先，拥有超过200个专利，成功打破了金融行业数据库国产化替换的困境，守护了国家数据安全。
TencentOS生态开放的国产操作系统
，与合作伙伴共同构建了“国产芯片+TencentOS Server+应用”的全栈国产化产业生态。技术自主可控，拥有千万级节点的应用规模，确保了技术的持续创新与安全可控。

互联网3.0操作系统ChainWeaver与腾讯云产品紧密结合，共同打造了一个支持国产芯片的全栈国产化产业生态，促进了软硬件协同创新，不仅推动了国产化替换进程，还彰显了腾讯在履行社会责任和践行科技向善方面的努力。

2.6 域名安全刻不容缓，急需打造一整套的极端情况预案

分类	当前问题
域名管理	内部重要域名不在MarkMonitor, 没有纳入知产部管理体系
	共用域名的情况泛滥，故障域过大
	缺少域名备份
可观测	缺失域名的DNS服务器变化的监控
域名故障容灾切换	域名故障没有容灾切换方案，止损时间会比较长
监管方建联应急	与海外注册局、ICANN建联、应急响应问题

方案

- **域名管理治理**
  - 内部重要域名纳入知产部管理体系
  - 共用域名拆分、备份
- **可观测**
  - 重要域名支持DNS服务器变更监控
- **域名故障容灾切换**
  - 域名故障容灾切换
- **监管方建联应急**
  - 与海外注册局、ICANN建联、应急

3. 未来规划和建议

1. 守住大盘。超大客户的IDC上云，各行业头部winback
2. GPU突破。增加在GPU上的服务能力，发挥软件价值
3. 出海突破。海外目前还处在较为良性的市场竞争环境当中

4. AI提效。通过AI提高效率，降低成本
5. 市场竞争。坚定信心，逐渐获得更大的市场认可
6. 质量保障。系统性提升产品稳定性，防范黑天鹅事件发生
7. 一些发展趋势。

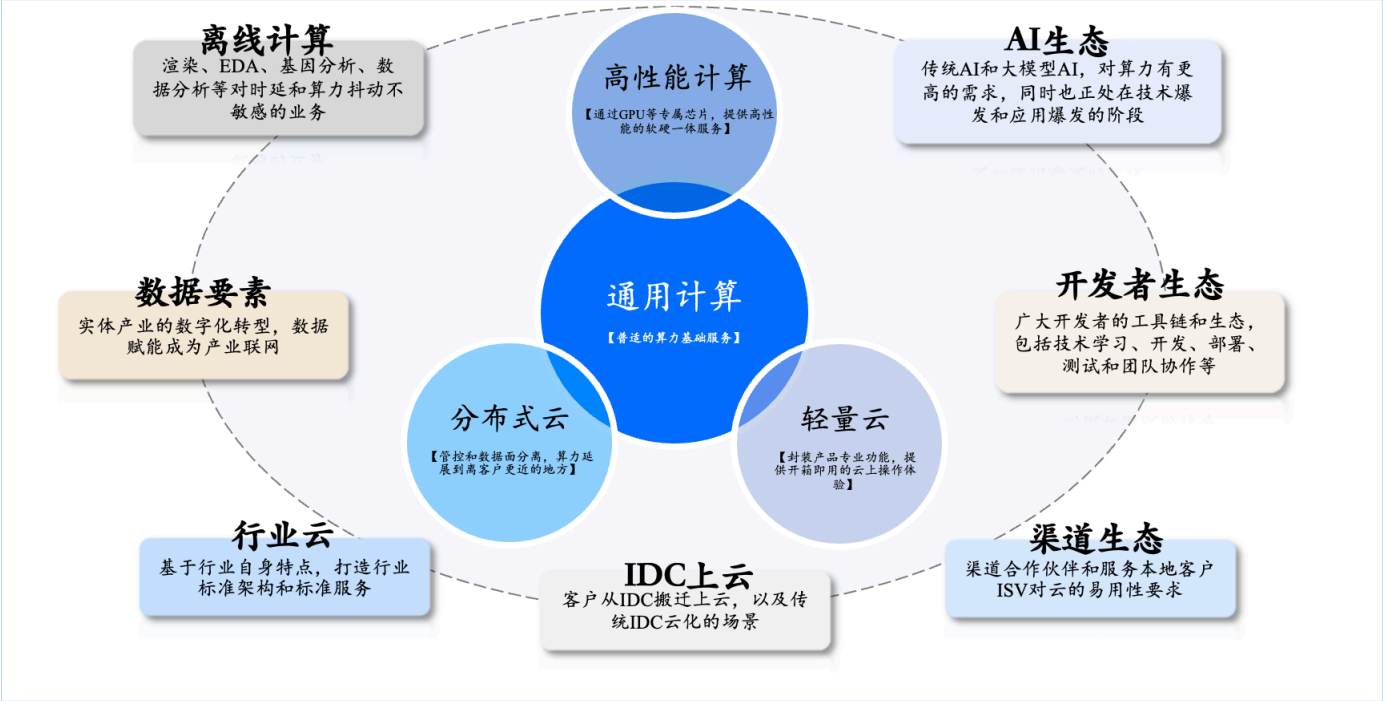
3.1 大盘稳健前行，经营结构已经非常健康

在当前的调度能力优势下，我们处在一个很好的状态，即使是在市场高度竞争的情况下，我们仍然有很大机会获得较大增长。

国内接受云的比例事实上还不高，但这是一个不可阻挡的趋势，进展缓慢的本质是中国的工程师红利，AI加速了云的势能。

3.2 计算产品的角度看，我们离客户仍然很“远”

1. 产品功能上的“远”。云计算替代传统架构最重要是对传统架构的兼容，但如果只有兼容性则永远无法发挥出云在计算能力自下而上的产品化抽象过程的价值，提倡“以云的方式使用云”，即让用户逐渐脱离对底层概念的依赖，使用高度产品化的云上组件，比如大数据套件、PaaS、SaaS、中间件等。宽泛来说，这是“云原生”的产品方向。
2. 物理空间上的“远”。公有云原本最基本的概念“区域(region)”和“可用区(avaliable zone)”几乎完美地匹配了大多数互联网应用的数据组织和容灾需求，这是一种以聚焦为主、分布为辅的模式。但是对于更广大的产业互联网客户来说，不论是数据存放的行业合规规范，还是实体资源分散的业务形态，均对云的分散部署接入有更大的挑战。通过将控制平面和数据平面的分离，将云建设到用户需要的地方去，是“分布式云”的产品方向。
3. 算力需求上的“远”。数据分析和人工智能的流行，带来了对通用算力能力的巨大挑战。通过GPU等异构芯片为特定计算场景加速，极大地提升了高性能 计算场景的效率，但是也同时带来了一系列需要解决的问题，比如与大规模计算/存储/网络资源的组织分配、ai模型软件栈的深度适配等。这是“高性能计算”的产品方向。
4. 操作体验上的“远”。云计算的发展，需要兼顾稳定性和灵活性，但在操作体验上反而变得越来越复杂，提高了中小用户的使用门槛，这部分客户包括了中小企业的开发者和个人开发者，以及对技术了解相对较少的技术爱好者和应用托管者，也许这其中就有数年后的行业巨头。在产品层面通过裁剪和包装产品概念，以“开箱即用”的理念打造轻量的云产品集合。这是“轻量云”的产品方向。



3.3 计算能力自下而上的产品发展

巴别塔：
云计算的本质，通往天国

西西弗斯：
不断的尝试，而似乎注定失败（弹性计算是否就成功了呢？）

共识不可能一蹴而就，与空中楼阁不一样，自下而上是一个渐进且稳妥地路线

3.4 GPU要把整个AI Infra结合好，提供又便宜又稳定的服务

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。

3.5 海外具有非常大的市场空间，是我们一定要去重点投入的方向

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。

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。

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。

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。

3.6 产品出海除了资源挑战以外，文档、体验和生态均有缺失

核心在于产品化程度不够

功能大类		详细描述
产品竞争力	国内站和国际站能力对齐	针对国内站和国外站的能力做全面的梳理，并排期规划补全差异，无法补全部分则以文档形式清晰呈现差异
	白名单治理	Review并全量治理白名单，对于已上架超过半年的白名单实现全面下架，并对新增白名单进行生命周期管理
用户体验	API 能力对齐友商	By API接口梳理，从产品能力和使用体验两个维度展开与AWS/GCP/Azure的差异化分析，并规划补全差异
	IAC工具生态适配	站在客户的角度体验terraform、packer等IAC工具，完善能力对标，并保障能力的持续更新，以实现AWS GCP Azure计算相关的API的无缝迁移作为目标
	文档描述深度优化	参考AWS的接口文档，全面深度优化产品文档描述，提升文档可读性和可操作性，实现100%可操作成功指引
服务拓客	售前材料/场景化方案沉淀	售前材料从对比阿里云到对比AWS/GCP/Azure等海外云厂商的方向进行改造，客户案例优选海外local。关注海外主攻的游戏、电商、金融等场景，by赛道做场景化分析，并输出针对性方案联动前端行业一起进行售前推广

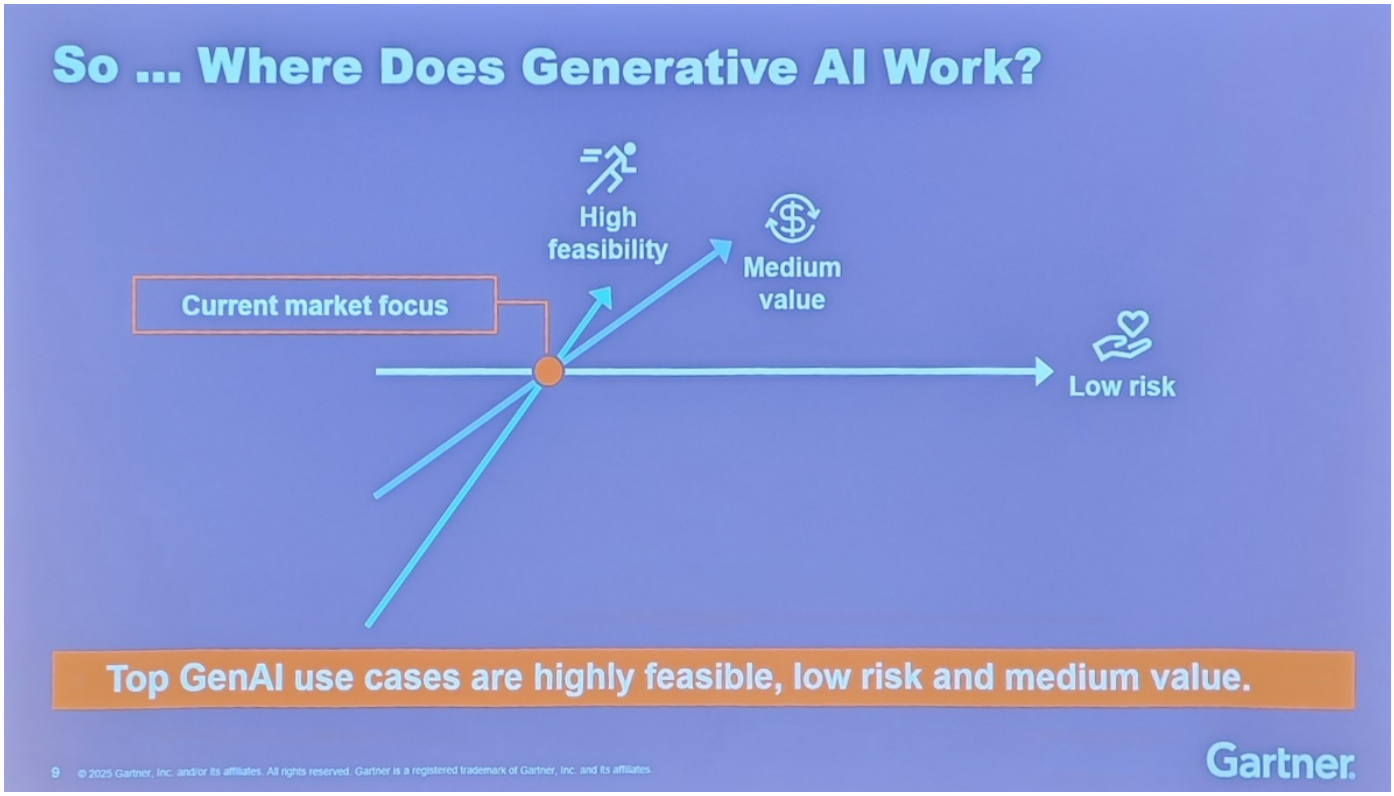
3.7 AI提效可以减少很多重复劳动，让团队效率提升

1. Coding提升代码效率
2. AI客服：CVM AI助手、运营助手、轻量云AI助手、OrcaTerm助手
3. MCP托管：为客户提供MCP执行环境
4. HAI推理加速
5. 文档智能优化

3.8 AI应用目前仍然存在很多问题，这可能会阻碍进一步发展

1. 效果不好预期，调试困难
2. 效果测评困难
3. 专有知识数据集建设困难
4. 仍然无法形成可有效指导的方案
5. RAG等技术本质上是退化方案，有悖于大众期待，且建设困难
6. 调研业界情况来看，大家普遍遇到类似问题，比如Google的轨迹跟踪其实是一种非常妥协的测评补充，反正证明了这里的难处

> Gartner分析目前五个典型场景
>
> - **营销定制化。** 例如可口可乐， 针对不同地域做营销。
> - **运营优化。** 例如财务，多模态提取数据，哪些是供应商付款等。
> - **客服。** 服务内部， 语音转文本、摘要、情绪分析。 黄金销售
> - **知识管理。**聊天窗口，知道内部、外部的信息，快速检索、决策。
> - **软件开发。** AI 开发，代码开发、记录、老旧代码现代化更新。 进展最大。 中美走在前面。
>
> **共同点：**
>
> - **客户可见价值， 不一定是翻天覆地的变化。并且可以说出来的价值。**
> - **技术可行性， 提示词工程，rag 这块的 SaaS很快结合。**



正则表达式与AI的关系?

正则表达式（Regex）作为一种强大的文本模式匹配工具，在计算机科学中已有数十年应用历史，其核心价值在于对结构化或半结构化文本的精确提取与验证。但随着AI技术的突破，Regex的边界正被大幅扩展。

场景	传统Regex痛点	AI融合方案	价值增幅
金融反欺诈	诈骗话术每月变异15%	NLP发现新话术 → 自动生成Regex规则	规则更新效率提升90%
医疗病历解析	“腹部疼痛”有68种表述变体	BERT识别症状实体 + Regex提取量化值	信息抽取准确率↑35%
物联网日志诊断	设备错误码组合超1000种	LSTM预测故障模式 → 生成诊断Regex	平均修复时间缩短60%

按照这个思路下去，还有以下几个ai适合的场景：

1. 政策解读
2. 法律解读，合同审查
3. 医疗病历理解
4. 跨模态工单理解

2025 年成熟的 LLM 应用均遵循 “专业纵深 + 实时闭环 + 成本锐减” 的三角定律：

专业纵深：放弃通用模型幻想，深耕行业 know-how 与领域数据
实时闭环：从“问答”升级为“感知-决策-执行”工作流（如工业工单系统）
成本锐减：模型蒸馏 + 量化技术 + 合成数据，使企业级部署成本下降 10 倍
正如微软 Azure AI 负责人所述：“2025 年 LLM 的竞争已从技术竞赛转向 场景渗透率竞赛——谁更懂业务，谁就能让 AI 真正创造现金流。”

3.9 市场形势发生了变化，各大友商思路不同

阿里云裹足不前，丧失了技术领先优势且已无挣扎意愿，策略上几乎all in AI，以不合规的GPU来冲大收入，可持续性不足。

华为云未从传统行业突围，技术上的缺失没有补齐，注重商务能力和国产化叙事，与互联网公有云厂商渐行渐远。

火山云低价策略效果很好，技术投入大，但是在业务方向上all in AI，收入增长快速的同时降低了结构健康性。

3.10 海外友商在海外仍有很大优势

AWS GCP Azure一直与国内的路线不一样，走的是高度产品化的方向。这是我们目前最大的问题。我们应当持续建设产品能力，同时以高性价比和良好服务能力拓展重点区域。

3.11 质量是云的生命线

通过产品化的能力将云的使用实践真正开始超越传统架构，提供安全可靠的基础设施

3.12 不同形式的混合云架构将成为一种趋势

1. 多个云统一管理
2. 公有云和私有云统一管理
3. 云和本地机房的统一管理
4. 本地机房获得云的延展能力

3.13 AI进一步促进云的开发模式和交互模式发生变革

以云原生的概念为例，云最终要提供一整套更现代化的IT基础设施。云的开发模式和交互模式受限于对传统的兼容，有很多技术和体验仍然是非常陈旧的，AI正好有机会促进这些变革：

1. OrcaTerm解决云计算体验割裂的“最后一公里”：带内带外的云账号打通

2. TAT助手可以全身云的Open API处理操作系统内的任意资源
3. 控制台的MCP化可以帮助用户快速理解和使用云的能力

3.14 端到端加密的数据保护

模仿Apple PCC框架，提供整个云上数据保护的机制，彻底解决用户对云的信任问题。

3.15 绿色能源与云的强大调度能力结合，同时调度算力和电力

云可以通过调度能力，提供差异化的算力服务，满足包括在线、离线业务、边缘推理、冷备存储、CDN等不同场景的算力需求。

3.16 数据的治理能力决定了AI、区块链、数据要素、产业互联网等诸多理念的落地

现在大部分的技术阻碍都指向了数据的问题：

1. 数据不存在：数据未定义，未与真实实体连结
2. 数据定义不明确：没有准确的数据定义，导致数据含义无法理解，更无法被代码和AI使用
3. 数据无结构化管理：以流式方式存放在各个地方，无法真正调度起来
4. 数据的安全性忧虑：缺乏数据保护的机制，从而进一步降低数据治理的意愿
5. 数据的价值难以衡量：由于数据难以使用，导致价值难以衡量

未来也许这会是另外一个需要系统性重构的领域。

4. 有什么问题

1. 出海的规划问题
2. 集团与子公司的管理
3. 如何增强内部协作