算力租赁行业调研

会议原文

你好，可以听到。可以听到。

那个，那个，您好，我是弘则研究的那个冯冠儒，然后之前我同事帮我跟您约这个时间跟您聊一下。

哦哦，明白，您说。

是。对，然后那个主要还是那个想关心一下现在这个整个国内的这个算力建设的这个情况，或者说因为现在讲的比较多的是这个算力租赁这个事儿，所以。所以就想看一下那个现在整体的这个状态，因为前段时间也一直在涨价，然后可能我看最近价格是不是有点稳住了，然后就想看一下后面的这个情况。我看见那个您那个同事姓姓罗是吧？对，他给我当时列了个提纲，先这样吧，就是按照您说的吧。就是现在国内的算力，就是前一阵，是因为这个，叫龙芯呀，像这个龙芯呀，还有一些那个智能体的应用爆火之后，打通了这个 C 端的这个 C 端的这个算力的这个，就稳定性的收入吧，就市场。

然后，原来都是做 B 端，好多好多那种机房呀，就是面对大模型客户，还有一些什么，互联网厂商呀，政府。

政府这些这种比较正向的应用投 C 比较多，投 B 端比较多，然后就一直没有爆火，有一些单了，或者大模型公司做做训练之后没钱了之后资金流断了，就整个的这些项目就断了，有这个中断期，就一直应该是不温不火吧。当然就是受制于国内这个卡的这种英伟达的高端显卡的这一直在依赖的制裁，整个算力还是缺的。但是这个随着龙芯这边的爆火之后的话，大概从今年过完年之后到这个到现在吧，整个算力市场的这个高端卡包括这种咱说低端卡、推理卡，整个市场价大概涨了 30%左右。

这个价现而现在高端卡是越来越缺，这个现货市场就是比如说咱那个像那个英伟达的这个 B100 吧，现货市场已经被炒到了这个。以这个超威为为这个超威 B300 为这个，举例吧。现在已经是，两 T 的是 640 万到 660 万，三 T 的话可能就是，已经 680 万到这个，680 万左右吧，四 T 的话已经到 700 万了。整个这个所以说这个整个的算算力租赁的高端市场，包括传统的这个低端市场，就是中低，就是比如说这个 5090 这些，整个的这些包括 A800，A800，A1100，A1200 这些终端市场现在都是一卡难求。

现在这个整个国内的这个租赁的这个租赁这个市场非常火非常火，就是没有卡。你现在别想找一个，比如说你想租个一年的三十二台的 H100，H200，甚至 A800，都没有，都没有，对。

这个比方说您刚刚提到那个过年后涨了 30%，这个指的是算力服务器的采购价格还是算力租赁的价格？

租赁价格呀，这个租赁价格是随着这个采购价格抬起来的。比如说，举个例子，年前。年前的时候，一张这个 H100 的这个是一个 H800 的服务器吧，H100 的服务器吧，当然说比较主流的这个，现在过年的时候大概是个 5 万到 5 万之间，那现在的话基本上是 6 万到 6 万，而且还找不着货。基本上就这，这就是这个，而那个，像 H200 的话，就从 6 万多吧，6 万出头，现在涨到 8 万块钱了，也是没有货，就是租赁市场。当然了，这个采购价格，像 H100、H200，以 H200 举例吧，然后，去年的大概是个内存涨价之前，九月份内存往上涨价之前。

这个 H200 的话大概是 220 到 230 万吧，不同的品牌，甚至有的跌到 210 了。但是现在的话已经涨到了这个，已经涨到了这个多少呀？就是像超微还是超微的话，350 万了。当然跟那内存涨价也有关系，但整个的这个，整个的这个就是说这个，叫服务器溢价吧，从期货到期期货到现货，包括这 B300，这个溢价的话现在有溢价应该是 30%到 40%吧。我举个简单的例子，就是去年。去年的去年的十月份左右，一台 B300，十一月份吧，B300 刚出来，一台 B300 的价格大概是个期货价格，大概是个就是满配的四 T 的，大概是个 460 万，450 万到 460 万吧，就期货，不知道是不是现货，然后，但是，现在那个现在那个期货价格大概就是涨要到 520 万了，就涨了五六十万，这是原厂出的，然后到这个现货市场，就是利润非常高了，到了你像到 680 万，就说里边有 200 万的利润。

有 200 万的利润，就是整个的这个，对，我再我再跟您确认一下，就是对 H100 现在的租赁价格是六点七到六点八万，是吧？六点七到六点八，现在不是价钱的问题。没事，我先确认一下价钱，然后是的，是的，是的，是 200 是多少？即使 200，现在在 8 万块钱往上。8 万，然后现在还有啥吗？现在还有 H800 吗？H800 有，H800 的租赁价格还是，因为 H800 也算是一个 10%的主流，就三个主流机型，H800 现在的价格差不多也得个六万三四吧，六万三四。喂？

我在听，我在听，那个 H20？

H 二零的价格相对来说的话，低一点吧。现在 H 二零，141G 九十六 G，幺四幺吧，幺四幺的价格差不多，在一个。4 万左右吧，应该是 4 万左右，4 万多一点儿。

明白，明白。这个比方说，这个价格层面上，它是比方说这个周度会变化，还是说可能月度变？

这个不存在周租吧，像我这边说的都是这种，咱说了，像八卡整机吧，咱说了就是那种比较高端的机型。因为我主要是也是做这些东西，这边一般是就是按照年最低租，最晚就是最少租一年，一年的合约，一年的合约的话价格稍高一点，比如说是个 6 万，就说这个是一百，就是 60013 万，你要是三年的话就能便宜点，六万四五，那么要是你要是租五年闭口的话，可能就是 6 万多一点，6 万块钱吧，6 万。喂，明白，明白，我在听。不好意思，我在这球场边上。

哦，您说。

我刚刚说那个，我刚刚说那个价格变化是指说，比方说我上周要是跟你签一年，和我这周要跟你签一年，这里边会有比较大的价格波动吗？

不会有比较大，但是就是水涨船高，就是每周你这个现在这个阶段吧，现在是还在还在一个就是资源资源这个紧缺的一个通道，而且是这个整机的价格上升的一个通道。那么现在你如果说晚一个月或者是晚一周来定的话，那肯定要这个市场是资源越来越少，你这边大家都来抢这个资源，都来抢这个资源来做这种算力的应用，包括这些，卖 Token，现在比较火的是卖 Token 这些，所有这些东西的话，你就是以抢到卡为目标，然后这个的话你下手越晚，这个市场溢价肯定越高呀。

明白，那这个比方说往后面去看的话，会有就是您觉得比方说拍脑袋，就是那个您觉得咱咱就比方说 H100 吧，现在是 676.76 点八的价格，您觉得就是后面还能往上涨吗？

我觉得会涨，我觉得会涨，因为现在就是以实际我遇到情况，因为我也做租赁，以前就是做租赁出身的，那后来就是转这个英伟达服务器的这种贸易，然后这个当然租赁的市场一直关注，现在这个租赁的话，这个是会越来越缺，越来越缺的，就是大家的话现现在有比如说我手头有十几个用户，有找 H100 的，有找 H200 的，都压。都压在我手里的单子压两三个月了，找不到合适的那种二一百二十八集群，什么六十四集群都没有，非常缺。现在市场上就有零散的，比如说十台八台那种小集群，小集群小集群在这儿在这儿还在这儿怎么说，就在这儿还能放出货来，就是那些倒下来的，有些用户训一点五训练完了，推理完了之后倒下来的，基本上就是满满的，现在根本上市场上就是可以说是一机难求吧，这样局面。

那比方说，我在听，我在听，那比方说那个，您觉得大概会涨到啥样一个状态？或者说现在咱跟客户聊下来，比方说他们能接受的这个价格的上限是多少？

您说是还是以 H100 举例吧，就是哦可以。H100 的话现在 H100 H200 吧，就是价格的上限。我觉得，要看这个，就是前一阵不是那个特朗普来了，要把 H200 放开，其实这些东西都是相关的。高端的卡像那个 B100 是禁的，什么国内，B800，B 300 B 300。然后，H200 说是要禁，说是要放开，那么全国内的十家的头部的企业，包括像联想，就是浪潮，还有这些什么字节、抖音、京东、阿里都获得了这个就是什么，但是现在并没有开始弄。这个特朗普来了之后，国家现在就是说已经国家现在已经那个。

国家已经那个就是说，表面上在开会的时候他同意了，实际上并没有放开，有好多的东西还是在交换，这个东西要传导过来，我只觉得至少三四个月，如果能真的放开的话。如果说真能放开的话，这个价格像这些整个的服务器的这个租赁价格，整个的上下游的租赁租赁价格都会往下降，都会降。都会往下降，我觉得会降 10%到 15%左右吧，一个大家就回来一个就是这个稳定值。

但是如果说 H200 不放的话，那么整个这个大盘会越来越紧。明白，就是您觉得您觉得如果放开的话，价格会往下走是吧？

会往下走，短期内会持平，然后。会往下走一部分，现在这个价格就是说没办法，没有卡进来之后，就被，就可以说被炒作了，也不能说炒作，就是说确实是这个情况。然后，这个全放开了会降 10%到 15%，也会很稳定，因为这大家这个也互相之间不会砸盘，互相之间不会砸盘，

明白，但这个怎么去理解这个事儿？假设说那个 H200，因为我看他那个新闻上说应该是放 70 万片吧，一共对吧？十家，然后。

十家对，一家大概是七点五吧，这个数字比较公平，七点五左右。反正那个联想是拿拿到了这个就是售卖权，因为他有这个生产能力，售卖。售卖权。

对，反正七十七十五万片，那对于现在国内的算力的总盘子是一个多大的影响？

75 万片，你除一个八吧，除一个八就是一个 10 万台不到的服务器，10 万台不到的 H200 服务器。那现在 10 万台。您稍等，我看大概是 10 万台吧，10 万台，10 万台的话。每台 10 万台，其实对整个的这个国内的这个市场的话，大概有个能占个百分之二三十吧，30%吧，三十到四十吧，是它是有一定的这个有一定的这个冲冲击的，冲击的，但是因为。说这个不能就是其实打开之后，打开之后的话，有些东西也改变不了了，也回不了原来的这个局面。原来是这个内存的价格比较低，硬盘内存这些东西。

现在是什么呀？是这些价格都涨上涨上来了，因为内存条一台整机里边内存的占比，原来就是占个 10%吧左右吧，现在都要占到 30%了，三十到四十了，一个内就内存。整个的

内内存硬硬盘的这个涨价，报就整机的价格就不会往下降，就不会往下降。就是虽然是那个什么就是数量进来了，但是这个成本上去了，这个价格也应该是不会下不会下来太多。明白，就是相当于国内存量的 30%，是吧？高老师。国内存量的 30%。这个但是，对这个市场会造成冲击，但是不会大量的降价，因为成本确实上去了。

你原来一台这个，一台就是说这个，在去年七八月份的时候吧，一台也就是二百四五十万，H200，二百二三十万，二百二三十万这个价格。你现在的话加上那个内存的涨价，现在已经冲到了二百七八十万的成本了。进进来之后的话，整个这个价格也下不来，也下不来。因为原来就是去年的大概是个七八月份的时候，H200，它就都有这个成本，成本、租价。然后，这个去年去就是去年的那个价格，那个价格再也回不到了，就是便宜也便宜不到哪，便宜也便宜不到哪里去。

明白，明白。

那个，我刚刚想问的是，就是因为我们这么算的话，就相当于供应会供给会多 30%。然后那假设说，假设说需求如果翻一倍的话，那其实这个紧缺会更严重，对吧？

因为你需求的增速比供应，您说的非常对，也就是现在这个 C 端市场打开之后吧，C 端市场打开就是说我吧，就说我们吧，我们以前都不用这些龙芯呀、豆包这些东西，就是没有这种习惯，就这个的话包括这些东西，拼命，但是现在就是说我们基本上就是说这个整个的这个 C 端市场会越来越成熟，很多人的生产工作呀，包括就跟当年用这个抖音，就用用这个用这个什么一样。就跟用这个微信一样，抖音一样，所有的这些东西，我们养成习惯之后的话，就基本上 C 端的。

B 端的市场是非常大的，非常大，越来越大。就是每天的，就是比如说调用这个，这个我们叫智能体，Token，调用这些东西，不管是专业的人员，还是咱说的咱这些生活呀，就是一些基本的这些这东西，咱们会越来会越来越大，对。

对，之所以就是说，我就想问这个事儿，就是那假设说需求端比这个供给端更快的话，那其实那不会降价，还得涨价，对吧？那十到十五块钱也不会降其实。

短期之内吧，会吧，短期之内会，但是，长期的话肯定还是涨，还是涨。明白。因为现在这个现在最牛逼的就是说，并不是说这个你单纯做算力了，现在好多大模型火了之后吧，就是每我们的这个，不仅是中国人，不仅是中国人，中国人这个自己的就是说这个用这个算力，日常的这些东西，智能体。现在，全世界的这种全世界的这些，比如说这个欧美的那些程序员，初创企业，他们，因为中国的电价便宜，然后，运维成本包括大模型，

咱们做的大大模型，像开源大模型，像这个像 Kimi，还有这个 MiniMax，是欧美的最贵的，像那个 ChatGPT 五，是它的大概是多少？

十分之一的这个叫叫 Token 的这个叫什么？我们叫十分之一的这个价格吧，就是。说输入 Token 输出 Token 还有这些都是它十分之一的价格，那么就这样了，全世界基本上都在用我们中国的百分之七八十，除了高端的那些，比如说像大厂，开美国的大厂科研机构他们用这个啥的 G P T 那些东西，然后好多的初创企业，还有这种我们叫独立开发者，都用中国的中国的这些算力大模型，那整个的这个全世界的这个咱说这些 Token 包括一些算力，中低端的算力，甚至一些就大模型的应用什么的都到中国来了，所以说整个这个市场是越来越大，越来越大，根本就不愁这个中国现在的拥拥有的这个算力服务器，这是个这是个标准化的东西，就是必须必须有这个东西你才能做算，做算力，做 Token 产生 Token 那么所有的这些东西的话就会造成中国这边的整个这个我说这就叫叫我们叫需求吧，是越来越是越来越厉害的。

对，就我这里边有个问题，就是之前也有人讲过这个就是所谓 Token 出海这个逻辑，但实际上但实际上我看像。国内的这些独立的模型厂，智谱、MiniMax，他们应该在海外也有云的供应，就他可以在海外直接买海外的云，就是用人家云的算力，就没必要说假设说我是个美国用户，我没必要就是我在美国提出一个需求，完事把这个需求回传到国内计算，完了再回传到美国，这个事是不现实的，对吧？

其实是这样，它的这个很多这种比如说编程，或者是一些这个咱说了高端的应用，它还是就是要求这个输出的准确性，或者说这个叫我们叫这个算力的这个叫什么专业性，它还是用美国的一些的，比如说像 ChatGPT，像这些就是这一些大模型是最贵的，也是最专业的。但是，大多数的初创企业，初创企业或者说的就是那个叫我们叫叫独立开发者，一些并不需要，还有一些应用并不需要就是做这个咱说那个用。用最专业的这个东西吧，用最专业的这个东西，所以这个，您说的就是我们，您说他们会不会，他们实际上是真的现在正在用中国的中国的 Token，包括东南亚的，我东南亚的量，我想，我想表达的意思是，我认可你的观点，就是我想表达的意思是，假设说美国的用户他用智谱，他在美国用智谱，他其实也是跑在比方说美国的云上面吧。

真正的计算的发生位置不是在国内。

智谱是中国大模型，那肯定是在国内，最多是在这个咱说了，在这个东南亚部署一下，应该是在国内，是在国内。像中国的大模型也牵涉数据，你可以在我这算，但是一些敏感的数据，这就是比较专业的数据安全了，数据安全，就数据出海，这些东西是另一个维度了，另一个维度了，这个，但是现在实际的情况就是中国的 Token 正在这个什么被全

世界的，就是开发。开发者来使用，就怎么便宜？怎么便宜怎么来，明白。除了这个像亚马逊呀，这些比如说像这个像 Meta 这些就说了，这些他们自用，还有一些高端的这种，咱们说大模型的这个训练，用他们来说，因为有时候准确准确的话，反而就是说这种意味着低成本，只要训练一次就 OK 了。

那么现在基本上那种别的中低端的，甚至说大多数的这些东西的话都。都是那个什么，都是那个这个用用的用的中国的 Token，而且选择中国的模型，基本上都能解决。

这个我理解，这个我理解。然后比方说现在就是因为我们刚刚有个假设，就是说如果需求端更快的话，其实你这个价格应该会继续往上涨，而不是说可能短期有冲击，但还会继续往上涨。那比方说现在这个需求端的这个变化，就是从比方说我们那些在手的客户的角度上面，是能看到有很明显的变化吗？还是说啥样一个状态？

最明显的变化就是，再举个例子吧，举个例子，一是那个就是我说的，我现在想做算力的，就卖 Token，卖 Token，我也想做算力，不管是出海还是在国内，就是 Token，这一卡难求，你明白吗？我找不着这个 H100，H200 的这种，就是那个成规模的集群。我说明，我说明一个问题，我主要是想，我主要是想穿透到最终端的场景上去看，现在是啥场景消耗比较多，体量比较快。有几个场景吧？有几个场景，第一个场景，就是说咱说的短视频的这种生成，对吧？

短视频，大家现在抖音上好多的那个短短视频的生成，包括这个龙芯的这个叫叫龙芯的，就叫龙芯，小龙芯这个到 C 端了，就是那个小企业甚至家庭，甚至个人在这用，对吧？这个龙芯的费用，比如说这个，也分这个低端高端，基本上就现在一个月几十块钱，六十块钱，高端的话就是一个月就得就得几千块钱，龙芯这些东西。这些东西的话都是大量的使用大量的应用大量的应用，这是一个场景。再就是什么咱说了这个，医疗、工业、医疗、工业还有这个金融，这些实际的情况是这些比如说这个咱说的金金融公司或者银行什么的都在大量的使用这种，这个叫什么叫智能体吧，智能体，然后这个来来改变原来银行的这种靠人工的这种咱说了这种服务模式，然后还有这个像再就是这个，其实还是那种大厂大厂的这种不停的这个，就整个的互联网大厂，像阿里、百度、腾讯、字节，做公有云的，像国内做公有云的，像这个就以阿里为为主吧。

它现在就是说它把这个整个的这些这个我们叫算力服务器，通过自持的方式或者是这种租赁的方式，租赁的方式，它自己不持有总资产。然后这种方式的话，它把那个大模型，把大模型把一些云服务的能力就是加在上面，然后就是供应给全国的，供应给这个全国的各个行业，比如说政府呀，工业呀，还有这个所有的大模型公司，这种它的体量是，就是会国内的这几大的公有云，华为云，腾讯云、字节就是火山、字节，然后阿里，然

后。然后，这几家的话，大概是占用了国内的这个，就说高端显卡的使用使用率，使用率的大概有 60%到 6%，60%多吧，65%，整个这个市场，他们就是主力的消耗。

提供提供就是以前叫叫，以前叫什么，我们叫这个叫通用算力的公有云服务，现在叫什么，智算智算的公有云。他们是最大的其实消耗方，你说他们是客户也可以，因为他们确实是为这些持卡的这些就是资产方来来来真的掏钱来买这个东西，然后他们再加工分发，可以说最大的一个叫叫我们叫聚合分发平台来做这个来做这个，对，他们是最大客户。明白，其次国内的智慧城市呀，比方说你说你说，比方说我们自己手里的客户都是啥类型的客户？您说。

您说是我吗？对。哦，我这边的客户，像我的客户，就是阿里，一个阿里云，二是字节，三是那个像那个润泽呀，协创这些，这些都是从我们就是从我们这个手里边来买卡，来买卡，或者说他这边我可以找投资方来买我们的卡，投给他们，跟他们一直有合作，

明白，这是我的最大的客户。

然后，一些比较松散的客户，就是像那个像月之月之暗面，杰瑞星辰，Mini Max，还有这些。像这个，就是诸诸如此类吧，诸如此类，就是说大模型的客户也比较多，还有这个深势科技这些，好吧，这些。

明白，这个比方说我不知道，这个您会不会有感受？就是比方说假设说像去年国内的这些模型，就是可能还是处在训练的阶段，那可能去年的算力消耗我们讲大头，就百分之大几十可能都是在训练的这个场景上面。那比方说到今年的话，应该推理起来的话，比方说今年这个比例会是啥样的呀？

推现在就是已经这个数据已经非常清楚了，已经非常清楚了。这个训练和训练完了就是要往各个行业用，往这个垂直方向用。然后，现在推理的，就是需要的算力的规模，按中低端算，中低端算力为主了。这个虚拟推理的算力规模应该是这个训练的三倍，就比较准确，三到四倍。就是随随着这个大模型训练完了之后，越来越成熟的在这个各个行业垂直领域的就是推理的成熟运用，那么整大家用的越来越多，那么就越来越多，就是原来是没有东西可用，现在是越来越现在是有了有了好用的产品，比如说龙芯所谓的这个东西，大家都在用，整个的这个整个这个训练的规就是推理的规模需要的算力是这个训练的三。

三到四倍，还是以后会越来越多，因为大模型的训练就是一个，就是一个把全全部的知识，全部的知识进行那个叫什么，就是我弄在一起，作为去做训练，都是智能体，对，对，去年的比例大概是多少？什么比例？

去年推理和训练的比例。

去年推理和训练比例大概是个一比一点五到一比二，往后推理的比例会越来越高，因为训练就训练完了，你知道吗？除非在特定的领域再做一些打磨，想超等一下。

现在全全，您说。稍等一下。稍等一下。您说推理和训练一比一点二是吧？一比一点五到二。哦哦，那个是推理比训练是吧？

训练比推理，推理会越干越高。哦，去年就已经是推理高是吧？去去年就不错了，去年有出了这个像这个 Deepfake、V 三什么的，这些当然是投 C 端的比较多，但是。但是，就是还有这种咱说短视频的，像这个抖音、字节这些应用，它那些所谓推理，像做视频什么的都是推理，都是推理，它并不是训练。明白，喂，我就是跟您确认一下，我就是跟您确认一下，去年相当于推理是训练的一点五到两倍，然后今年变成三倍，对吧？大概是这么，现在是三倍，就是说整个的就咱说了这个推理算力，就是中低端的算力越来越需要的越来越高。

越来越高，明白，明白，所以就还是比较明确的，就是我们刚刚说的，就是需求端肯定是更快的，这个事儿肯定是没错的，会越涨越快，我觉得会越涨越快，越涨越快。明白，明白，然后那个刚刚您提到就是在几个需求场景里边，可能现在互联网大厂的消耗可能是 60%，是吧？

60%甚至更多呀，他们现在就是把那些中国的咱说了，从 H800、A800、H100、H200，然后 B200 很少，就是一直没有大规模用，B300 这些东西全都是他们，就是高端的卡基本上就被他们给吃尽了。然后，流到市面上的这些卡会带一些像就是比如说这个我们说政府、银行的金融的项目上，还有军方的项目上，他们这个对这个市场的价格就是溢价什么的不太敏感，有钱，对吧？有钱，然后他们这边大概吃了个 20%左右。我再跟您确认一下，我在。

我再跟您确认一下，您说的 60%指的是算力规模的占比是吗？

就是卡呀，你可以直可以直接认为卡，就是那个整机。整机，对，就是高端的这个算算力整机，算力卡。60%，六十吧，都在这大厂。

我为啥要问这个问题？是因为算力规模不等于算力消耗，对吧？

因为你有的。对，对。好。但是咱说的是，就不用说是什么概念了，就是咱说说的实际情况。大厂拥有这个高端算力的，全国的 60%。其次，像这些政府呀，一些银行，还有军方拥有这个 20%。甚至阿里他们，不是，甚甚至中国最大的这个，我说了这个干政府项目，智慧城市，信创这个项目的，华为，华为这边是中国最大的就是买。就是买阿里英伟达的这个叫模组，可以说持卡方？为什么？他们 10%的一个项目，信创项目，比如说一个

亿、两个亿、三个亿、十个亿吧，有一个亿用他自己的产品，阿特拉斯、昇腾、910、麒麟、B、九幺零、麒麟。

然后，剩下的90%全都是他也买英伟达的卡，让那个谁超级变超级变给他加工成这个整机，对吧？

就这样的，然后假设说，假设说我们从真正的算力消耗的角度上去看，就是真正开始工作的这些，真正产出的 Token 的角度上去看的话，我们刚刚提到的几个场景，短视频生成，龙芯，然后一些行业级的应用，然后或者说互联网自己，那这几个场景如果区分开的话，就是从算力消耗的角度上大概能占到多少？算力消耗，我是完了还忘了一个场景，最普通的场景就是编程，编程这些，这些。

互联网大厂占60%，然后，初创企业就是那种小规模的，用这种他不愿意用大厂的算力，用用这种咱说了，好多的数据中心，小数据中心或者小的这个六十四台、三十二台这种规模的也有很多，那么他们这边的话大概还占一个15%左右，再就是，像地方政府、金融机构，像就像这种这个做这个信创新创的这个项目，也是智算项目，信创的这种项目，大概占了20%到三十吧，就这么一个构成，这么一个构成。这个不还是算力规模吗？

那您说的是我一直没搞清楚，你这边说这个算力租赁的客户是什么意思？就是我想了解的是真正从使用的角度上面去看一些主要的场景是啥？比方说你有十台服务器，但你跑不起来，那不就相当于没用。我是说真正跑起来之后，发现场景比较多是啥？

那我现在跟您说吧，就是说这个现在根本就不存在这个，咱们说了一个算力的使用，我们叫使叫叫叫使用率吧，使用率，使用率现在整个算力的使用率在中国吧，不管是高端计算算力，现在基本上是在一个90%到95%之间，没有现在实际上没有这种就是空余的算力了，不用的。那比方说政府信创这种客户也都能跑满。他们的，我们是我，我认为他们政府信创这种项目的话，不是跑满是占满，他们是断开的，也不对外用，他们自己好多信创项目用起来之后，根本就不跑，你知道吗？

国产算力就基本上就可以说是这个国产算力的话，你就暂时不用考虑吧，暂时不用考虑吧，国产算力，他们的使用非常的低。

所以说这个说的是市面上的英伟达卡的使用率，90%到九十五至少。

对，我就说不管是不管是低端高端的全都这样。真正的国产算力的使用率非常低，什么样？因为它这边价格本身价格高，你知道吗？信创项目价格高。我提供同样的算力，英伟达的算力跟这个比如说这个华为的算力，它九幺零比九幺零 T，价格还有海光、阿特拉斯、海光那些，它整个的价格大概是英伟达的这个一台服务器八卡吧，它大概是一点五到

一点到两倍吧，硬件价格。然后还有半年的就是说这个系统它要我们叫适配，就说像库大这边大家都在库大上变成在这边跑大模型，一个任务马上买来就能部署，编程也是它的，也是这个叫英伟达的这个叫库大这个生态，就是直接可以用。

而那个你要是用这个用。

对，还是回到那个我们还是回到推理的场景上面，因为刚刚提到就是推理的算力规模，今年相当于要扩一倍，对吧？就是你要干到训练的三到四倍，然后这里边推理的主要场景就是我们刚刚讲的编程、短视频生成、龙芯呀，然后行业应用对吧？这些，就这些如果去区分一下的话，现在哪些是最快的呀？或者哪些占比最多？完事这个增速最快。龙芯，龙芯是最快，龙芯永远是最快的。

龙芯就是像以前的这个我们说的这个就微信，微信一样，抖音一样，从不用，但是用越来越多。

但是龙芯，但是龙芯我可不可以理解为就是因为你龙芯执行任务的时候也是在编程，所以这个其实穿透到最后其实还是编程的需求对吧？

还有这个咱咱再说了，这个各个智能体吧，像豆包，对吧？豆包这些。什么字节呀，这个就是自己的产品豆包，还有这个什么，百度、阿里这些，像什么通义写，就是通通义千万，这些什么全都是这些东西，这些东西的话都是，都是咱说是，都是非常越来越快的使用。当然我这边主用是比如说用豆包呀，用龙芯这些，别人这边，使用的情况我觉得也会，你用的越熟，就是说是可以说是越来越多吧，就是用的用。用的用的用的用的这个咱说的这个量是越来越多的。

而且也是会成会一个从不成熟，大家都不会用，到一个到越用越熟，它会是有一个一个叫叫叫什么？就是推理这个推理的这个什么指指数性的增长，对吧？

明白了，我就我跟您确认一下，就是您是相当于就是把把这个服务器什么的卖给他们之后您就不管了是吧？就是运营，什么运维，就运营这些您负责吗？

当然做了做了，我们是那个我们是这个从这个从这个就是从英伟达的这个代工厂下单，帮助他那个就是订服务器，到这个运回国内，到那个给他这个压测完了之后，给他上架，给他上架，我们叫交付实施，然后，包括后面的这个叫我们叫运营，运运维运营，都可以来给他做，因为这个我们最早做的是，比方比方说您能看到就是现在不同场景的这个后台的消耗的流量吗？这个很难，这个很难，这就是这个，比如说我们给阿里供了这个幺零二四台的这个单了，二零 23 年二零 24 年的七月份，那是 H800，我们给他定了，他们给我们杀了幺零二四，我们只供了 384，然后就派了大概有二十人的团队吧，在广州，广东

的那个是东莞普洛斯机房，给他那个做这个做维护吧，做维护，做维护之后，你说你说他后台的数据我们很难看到，但是这边的话像这个像这个什么呀，他的这个客户来访的时候，比如来访的时候有些问题需要解决的时候，我们都是配合的。

大模型公司，就是金融机构，甚至包括这个咱说了就军方好多，好多好多的项目，都是来我们共同的，因为处理问题，软硬件，我们都要配合。但是你说他具体的一些数据，后台的数据我肯定看不到，这个东西违规了，明白？好吧？明白。明白。所以我的理解就是您只能定性的去去说，反正需求肯定是不错的，但是比方说需求到底多快，这个其实也很难也很难给个数对吧？其实也是一直关注，一直关注这些东西，关注。但是你说这个准确的数数据吧，这些东西太难了。

这个东西你我觉得就是您今天也突然提提提这些问题，提这个问题，就是那个这个东西的话，你当然需要专门的调查了，对吧？比如说这个国内的这个，看您能不能接触到。就是可以接触到这个东西要专门的我去再找我的圈儿，没事，就是我们自己的圈儿，我再我再做专门的调查。您如果说单独的让我现在这个时间就给你一个准确的，那等于我就不负责任了，那属于是那属于不是说瞎蒙，大概数能说过来。

对，没事，那反正。那反正我听，我听您的状态，感觉对这个事儿还挺兴奋的，对吧？对，肯定是。

我对，对，对，是因为现现在，就是比如说我吧，我原来在青岛，就是全国各地跑哪里有项目呀，有客户，现在基本上我就要住常住深圳了，下个月去常住深圳了。他那边专门设立的这个，比如贸易公司呀，或者什么深那个 Token 分发的公司，要成立一个做 Token 分发，因为这个业务太火了。我们叫我们叫我们就相当于学习，你不能说我们是阿里，阿里他们是大的公有云，对吧？我们这边的话就是做做一个小的这种 Token 的分发。

就是大大大模型，比方说自己有，对，那个比方说，比方说那个，因为您刚才也提到，就是如果到推理这个阶段，实际上对芯片的要求是没那么高的，因为你像 5090 什么的也可以用。5090，A800，对，都可以都可以。后面我们的想法就是，比方说对于国产卡的想法咋样？因为，反正听说什么昇腾，然后那个寒武纪，现在也都还可以。

希望国产卡能起来，但是现在实际的情况来说，你从这个国家的政策角度分析的话，他肯定支持国产卡。但是，从这个商业不讨论，商业我们要成本不讨论吧。对，不讨论不讨论政策，就只讨论说他们就直接在场景里面跑起来，横向比较到底咋样了。

非常差。

非常差是吧？我实际的情况是我知道的几家像无问芯穹这些，包括这些并行科技，这老板都跟我认识，原来买我的卡，现在他们这边也还是整个的这个他们的算力平台所谓的什么易构，国产也跟英伟达的易构什么的只是个噱头，只是个噱头，用起来的话他们 80%到九十还是会用英伟达的卡，性价比高。这个我，这个差是从什么角度上差？国产的算力的。叫咱们说是还是刚才我说的那几个，就是一是那个价格，二是生态，三是那个真正的技术，就是说这个咱们能力，能力各方面的话都，说是什么达到这个英伟达的百分之 H100 的，比如说达到你 H100 的百分之什么八十九十了的测试，实际上连 60%都达不到。

他们是他们真正的这些这几个做算力聚合分发平台的，他们做这个异构算力分发平台的，我刚才说就是说这个比如并行科技，还有这个叫无问芯穹，这马上要上市了这些，他们这边跟实际的他们的就是技术人员跟我们跟我们沟通的时候聊的非常详细，说是国产卡根本就没有市场，我们也不用。这个是项目，是我们项目，我这边给你给你。主要问题是比这不，是你说主要问题，主要问题是比方说掉卡率比较高，还是说你真正算经济账的时候，比方说算 Token/s 的这个产出效率没有英伟达好？

各种综合的，综合的，综合的，甚至就是生态，就是说这个叫叫我们叫叫软件的适配，所有的这一些都不行，都不行。就有个客户，我如果用国产卡的话，我是花了钱不合算。时间，编程要改编程半年，然后比的全都是成本，然后还有这个所谓的所谓的这个叫所谓的这个咱说的这个算力的，就是说产出，他用一辆电，他产他用的产出的算力不行，这不就本身就很过很咱说的就是很拉胯，对吧？我这个各各项成本这么高，价钱也投上了，这电用的电都一样，产出的算力不行，它就意味着什么？

就意味着只有一些国产的，国产的信创项目，它不在乎像军队、金融，它不在乎钱。那么国家让我用这个国产算力，那我就用呗，反正我有的是钱，对吧？那无所谓，但是只要牵扯商业应用，真正的按商业规则来的应用，这些。

国产卡就不行，那不就是，对，那你这个结论不就是像我说的那个，就是你算经济账，比方说 Token 产出的速度这些就没有英伟达好，所以算经济账算不过来，对吧？对，所以还有体验，就是咱说了，各种体验什么的都不行，明白，都不行，就是生态这个，所以说这个事儿，这个事儿。这个事儿，这个事儿，如果动态的去看的话，比方说今年比去年，或者说比 23 年、24 年的时候，这个差距是在变小还是在变大呀？

变大，差距还在变大，越来越大。你好像觉得那个什么国家在那追赶，但是实际的情况是越来越变，差距越来越大。因为英伟达发布的这个比如说 B300 还有 GB 系列，200GB300，现在再往下的这个入就是入品系统，入品系统它其实也是它在就是这种代差越代差越来越大。知道吗？他这边也是会，因为啥推出新的一代这个咱说的从这个 B200 到

B300，一是 200 到 B200 的话，大概是半年就是一对，我们就要叫迭代了，迭代了。而中国这个追赶真的不行，我说这个不是说咱不努力，而是说这个什么呀？

就是说咱咱的咱的短板在这个叫什么？我们叫光刻机，这第一步就不行。明白。这是个这是个硬条件，谁说也没有用，你除非自己现在不是咱们勉强用这个七纳米的光刻，就是那种旧的机器，从别人的地方买来那些旧的机器。但是现在美国这边是零点三纳米对吧？制成主制成，再马上再往下的话是零点二，零点二，就比。就不用说了，0.3 这边就是咱们的还没到 0.5，是吧？已经落后一代半了，而且整个的以后的这个落后会越会越来越多。你这他们到两纳米一纳米的话，那时候真的就是整个的这个咱的这边的功耗呀、集成度、训训练速度越越差越大，越差越大这块。

这块不是说哪是个单点努力，咱大模型咱们是大模型，咱们意思是混合专家大模型，可以做的说是这个训练的量和他们的就是四分之一到五分之一，就是省省成本，但是还是不行，你这个硬的条件确实是还早。

明白，明白，我再问一下，就是比方说现在那个回到那个算力租赁那一块去看的话，就是。就是涨价到底是一个，就是我听起来有点像是一个成本传导，就是因为你内存涨了，然后包括我们的定价逻辑上面也是因为我的这个采购变贵了，相当于我折旧变多了，所以我要涨价。就是所以如果是这种情况下面的话，其实我们自己就很难多挣钱吧。

您说很难多挣钱是啥意思？

就是比方说比方说现在看租赁的话，就租赁这块业务毛利率是在往上走吗？

你可以这么认为，就是，现在，我就一直我想跟您沟通一个事儿，就是说您是想要现在入场，还是您原来已经入场了，那么你现在的那个旧设备就有议价的这个，就我们叫叫议价的，议价的叫叫什么？已经非常议价了，就是你比如说用了三四年的旧设备，而且现在还可以多高价，知道吧？还可以往上涨。但如果你新入场的话，那么你要承担这个中国现在的这种禁运，就是英伟达高端那个中高端的这个显卡的禁运带来的这种价格的这种我们叫成本的，就是硬件成本的溢价，对吧？

包括这个内存，包括硬盘，包括所有的这些东西，都溢价了，对，那么。那这个事儿我们刚刚不是聊过，就是你后面的需求还挺好的，那你还能继续涨价，那其实我现在入场去干这个事儿的话，应该也很快就回本了，对吧？整个的这个回本周期吧，这个我也其实我也准给准备了，就准备回本周期大概是一个三到五年，而且你现在三年你只能按照五年，三年的话，如果说按照三年测算，那么。那么你会有一个问题，你三年的话，你这个

价格，你这个价格就要就回本，你要定的价格比别人高，那你在国内是没有市场的，对吧？

现在大家整整体的测算都是五年，就算溢价也是按照五年来，那你这五年的话，其实这个就是说这个，刚才说什么意思嘞？

刚才我有点走神，您再说一下。

对，我就想，我想说，就是我就核心关心一个问题，就是现在需求变好了，对吧？需求变好，然后那个租赁租赁涨价了，对吧？然后这个钱我们比比方说比去年的时候是不是多挣钱了？因为如果多挣钱了，当然多挣了，为什么现在这卡都买不着，买不着卡，现在真的是这样，就是不管是租赁，包括现在这现货市场，比如说有两个客户各要六十四台的这个一十八这个 B300 的现货，他现在找我买了一个半月，我拿钱在市场上都买不着，你明白意思吗？那我是主力是出卡的做期货的，但现货我帮他买都买不着，因为那货主都是从我这采购的，那个卖高价现在就整个的现在这个确实这个算力非常挣钱，大家愿意议价，议价的话在这种情况下也就会也就会努力的去买这种高端卡，六百六六百万七百万都无所谓买，买了之后租出去挣挣钱，而利润非常高，原来利润率大概也就是个十到二十。

就是大家内部卷卷的，现在基本上是 30%的利润，比如说一个 B300，现在就是五年期的 B300 十六。16 万到 19 万，原来就能卖个 12 万，12 万是吧？现在 16 万、19 万都不卖，就这样。就是那种 16 万、十九万在那抢，对，整个这个平台的这种，反正就是反正就是现在租赁涨的价格能抵消掉覆盖我们能覆盖这个，就是原设备或者原材料采购的这种溢价，都能覆盖。

明白。那比方说，比方说这个，咱还是拿 H100 来举例子呀。就是你现在是 676.76 点 8 万的价格，假设说掉到多少的时候是打平？

您说是，我觉得掉到 4 万会打平，4 万的话会基本上会打平吧。但是现在 H100 已经出了三年多了，已经出了三年将近三年了吧，H100。这三年它当时的进货价就是 210 万，现在它旧机器能还能卖 210 万，基本上白租了三年，就相当于这个成本没，就是它在没有折旧，可以这么认为。哦，明白。明白，对这个旧机器在中国也有价格，包括这个最早的 4090，包括什么 A100，A100 现在 A100，A100，A800 这些卡，这些整机都买不着，有的是人要没人没不是没人卖，而是说真没有了，市场上，你明白吗？

这些算力就让人家用作推理，用作推理训练都可以通用的这种，这种在市场上都，你现在想买个三台五台的 A A800，A800，A100 你都买不着。是现在是这样，现在 A800。所以说整个中国的这个算力租赁市场，不分咱说推理跟训练，都会持续的还是刚开始，这个

算力其实开始也是两年多，不到三年。随着渐渐的成熟，各种生态的这种我们叫生态的互相之间的这种支撑吧。训练带动推理，然后推理的话会可能也会对训练提高更多的要求，因为推理大家这个叫什么呀，推理的这个越来越成熟了，越来越专业了，可能需要新一波的这个训练。

就是什么样的参数，什么样的这个我们说的行业的要求，会出现好多的那种这种咱们说比较通用大模型了，我们叫专用大模型。比如说在医疗，在生物，在一些在工业上，可能像一开始国内就好多大模型就不大一样，像紫东太初，紫东太初就是做工业做医疗的，就是工就是叫什么院，就是这。这个叫叫叫叫叫什么院来，中科院对吧？当时我们最早的租赁客户，就大家的话这些东西，我觉得是整个这个市场是比较方兴未艾了，刚开始，绝对是个蓝海，还早，明白，这个比方说您跟身边的这个小伙伴们去聊这个事儿的时候，大家态度都比较像是吧？

都觉得还挺火热的现在。好，这现在大家好多原来就是不大感兴趣的，现在基本上，基本上应该是当成一个热点了吧，热点，不管来问我想做贸易，想做租赁分发，甚至就是想就是要做 AI 创业，好多好多。现在好多的就是，要是说没有什么想做技术方面的东西的话，那么他们现在比较感兴趣的就是这种 Token 的这种，Token Token 的生意，咱不出海，咱不用说出海，就 Token 在中国，你如果做一个叫聚，叫叫我们叫聚合分发平台，非常。你这边就把这个，先不说全世界，咱们行了，因为你做图片分发，只能用你的，你自己买硬件，高端低端的硬件，对吧？

为了控制成本，这第一，你买，第二，你找机房把服务器布上，第三步的话，那你就把这些开源大模型，你可以跟他们谈，他们。他们都会答应给你这个叫，给你开源的，就是给你开源，就是授给你授权，API 给你接上，然后你可以就是往全世界，然后是往，像国内不说了，可以往这个全世界的这个我们叫，叫算力的使用者，对吧？不管是高端低端，就是通过这个大模型，对吧？就是直接就。分就是分发给他们，你可以中间大概挣个百分之，有挣平台的话能挣个几，比如说比 2%、5%。

然后的话就相当于一个淘宝平台，那你上边的从你这买 Token 的买 Token 的用户，对吧？直直接去他们往外卖，他们的利润可能是百分之二十三十，甚至甚至翻番，卖给欧美的话就翻番。我的 Token 确实便宜，是他们的 Token 的大概是五分之一到十分之一吧。明白，然后？

行，那我今天应该就是这些问题，反正那个定期跟您跟踪一下这个产业里边的情况，最起码感觉现在需求上没啥问题，然后供给还在持续短缺，价格上面就扛得住。

还会涨，还会涨，这个要看中国的这个 H20，它到底是不是真的让放，中国表面上是答应的，但是，明白，有很多东西都是业务交换，都是交换，如果特朗普不把一些东西给咱给咱放开，他不会，我答应你了，开会大宴会上答应你，真正给你的话，是不是你要拿拿出那个对等的条件来，对吧？这个东西还有个过程，我估计三个月的，三两到三个月的要过渡期吧，即使答应了他们也马马上过来不了，那个库房里也没那么多货在那在那堆着呀，还要生产。二十对吧？

明白。行，反正定期跟您聊一聊，谢谢您时间今天。

好好好，谢谢。好好，再见。拜拜。拜拜。拜拜。