### 20210709\_134108.m4a

说话人1 00:00  
一个讨论组，然后我们有三位主持人，特别是雪峰总，然后晋辉，金辉总还有唐总唐国，然后我们还有在开是你要不你说你在我们组不说了哈，你们得拆开对吧？你们俩得拆开。

然后先请雪峰，还有静薇，还有我们军先来做10分钟左右10分左右一个一个引导发言，就是从框架和编译器本身就作为从业者的专家，看看我们自己有什么样的问题，那么其实上面其实也听到了芯片和算法这边的input，所以也可以 depth的去加进思考。

那么结束之后，我们就会各个讨论的主持人会有一定的侧重点去进行分组讨论，那么分组讨论除了主持人要适当的引导对吧？因为如果有的嘉宾可能聊天了对不对？

你可以引导得引导回来，我们最后要是要通过小组讨论的发散，要逐渐收敛，所以在每个小组内要去对一些结论进行收敛，比方说能有哪些结论是我们得出的小范围形成的共识，那么它就是结论，那么有哪些是问题，你可能觉得暂时讨论不清楚，那么可能把这个问题抛出来，在整个会场里面再到最后去共同去讨论，所以这是目前大概的这么一个形式，所以可能每个小组主持人除了要引导之外，可能还得再找一位嘉宾去做一下记录，你们到底形成什么结论，什么样的问题，基本上是。

说话人2 02:03  
这么一个顺序。

说话人1 02:08  
好吧，你们都是在公司里面都是大佬对吧？我们在这里大家都平等对不对？好，我们先有请金雪峰，来自华为的中软的金雪峰，他是my sport的主要的负责人，好，大家欢迎。

说话人3 02:31  
谢谢蒋老师啊。

今天很高兴，我感觉好像是一个大型的网友见面会，有好多朋友在微信包括网上都聊过，今天终于见到曾建武非常高兴，然后我今天主要把我在马斯报做的那一些感悟，反正给大家介绍一下，抛砖引玉。

说话人1 02:56  
国际内容有可能会超时，没关系，你们如果我们再自由。

说话人3 03:01  
第一个我想介绍一下整个AI编译器的一个概念，就是讲一下我对API编辑的一个理解。从我在知乎上也大概介绍了就是说说我认为就是说在耳边气整个的他的驱动力主要三块，第一块就是说开始因为我们整个的AI的框架，包括我们都是在Python里面，所以说实际上很大程度上我们要编辑干的事情其实在加速开始，这是一个。

其中第二个就是dsa是吧？如果说我底下的硬件都CPU估计我们第二编辑就跟普通编辑是差不多的。

第三个就是我们深度学习，包括VR领域的业务的一个驱动，然后我们看一下Python，因为我是大概在零几年的时候就开始玩Python，但是我们那个时候还没有什么深度学习是吧？

因为我做paiseh就是工具，做一个测试工具，但是我看了一下ade的框架，我发现其实要玩的跟我们以前玩过的都是一样的，我们就AI框架搞的事情我们以前都玩过，因为Python的话整个的对Python的加速总共有三种范式啊。

第一个就是说叫c pass，最主流的方式就是说我Python很灵活，但是我有些地方性能是不够的，怎么办？就加一个CC加加的库把它替换注册进去是吧？所以说整个的就是c拍成这条路线，其实发展已经很多年了，他的方式非常的粗暴，就把我内部的拍成解释器的内部的数据结构啊，结果就完全的开放给开发者，大概有1000多个，它这种形式这个形式我觉得支撑了整个拍摄的生态的发展。

但它有一个问题说。就是说他也阻止了他也是想接着编译器的一个眼镜，大家看一下，就我们看第二条路线就拍拍，就拍拍他其实是希望提供一个Python的类似于Java的这种接口的编译器是吧？去搞定Python语言执行的性能，但他很难成功，为啥？因为 c派生把内部的东西全暴露出去以后，大量的 C的兼容性他搞不定，是吧？但到现在为止他离职兼容了20%不到的结果。这是第二条路线。

第三条路线就是那么把这东西假设，然后这是一种对拜城街的编辑的一种妥协，然后反正我是做了一个旁路是吧，我全部搞定，我就摘一些出来搞定。

然后我们可以看一下整个的 AI框架的饭是早期，我觉得他盗窃就是典型的CEA paiseh是吧？啥都其实就是一个事情还是，很简单，我的执行逻辑，我的很多语法就是依托拍摄，但是我性能不行的算是用 c对吧？然后它是follow1.0，我感觉是一种CP is跟 number的一种结合，对吧？

但是到后期我们都看到就所有的框架，包括开始它就是可以的，包括它的土豆价格是其实走的都是c派生加number的一种结合，因为number你们去看它是没有他的目标是没有c扩展，所有的Python的语句都可以落到我们的二进制区是吧？

但实际上这种情况在深度学习里面可能是不可行的，因为他的很多算子，比如说矩阵层这种性能可能很难搞定，用Python这种直接就按照这个二进制很难搞定，所以它最终就变成说我py可以用j的就是修饰方式来加速，然后性能高的算值，我还是用c的一个筛选去搞，这个是一个。

第二个驱动力就是dsa的硬件加速这一块，其实我觉得讨论很多，现在我觉得就是说中午也跟唐博在讨论，就是说整个的我们的硬件，就是说所有的都在做dsa的dsa化是吧？我们传统上比如NPC里面，我们有华为的酮，里面有VR的CPU单元，也有web的，也有标量，反正也有来举证是吧？

但是我们现在看到CPU，包括苹果的，包括英特尔的出来以后，现在也加了就是举证代表了，所以说这两个架构其实都是类似的，无非就是说在npo里面我举证单要多一点，CPU少一点，是 CPU可能还是为中心的是加速器是吧？

是，所以说我们在npo上遇到的便宜的问题，可能未来在CPU上也可能会遇到，这是一个问题，然后整个编译的话就是说他要解决的问题就是说，因为我们最早的像他掏钱这种框架，他的算子跟图层是分开的是吧？

然后蒜子的关于蒜子的，然后涂层是管涂层，但发现其实这种方式是没法充分发挥dsa架构的这种性能。

一方面说它的算子性能不是最优的，我还可以通过two name，就是说比厂家提供的算是更优的这种方式。第二个可能他的算式的划分就是。

说话人4 08:33  
算法工程师。

说话人3 08:35  
划分的算的力度可能跟我硬件能够发挥最高性能的算值力度是不一样的，所以说我可以做融合对吧？所以我们现在看到我们几个主流的编辑，像叉a它就是呃最早开始做图算融合的，然后tbm在叉a的基础上它是可以做自定义的是吧？

因为超a它基本上是这个规则确定的，然后后面像TC包括MS包在搞了一kg，我还希望就是说这种是概率，包括t0是自动化的，这是一系列的发展路径。

最后一个就是其实上午很多专家也提到了，就是说驱动arbh还有一个巨大的因素，就是我们整个深度学习，这种负载第一个比如说自动微分这个不介绍，还有刚才讲的大模型带来的并切分，包括我们一些量化，这个是推理包括常用的一些东西，这个是我对整个的ERP驱动力的一个理解。

然后整个PR边区的范畴，我想主要就是因为er BH是一个领域特定的一个多层的编辑，这一块估计大家也比较熟悉，主要是分三层，一个是图边区，然后三个边区跟周边建。

我们图片一期里面，其实我觉得大家有很多不同的定义，就是说图片一起有的人把前端的转换其实没放在编辑范畴里，因为我从拍摄拿过来以后，怎么把它转到我底层的我的图的 r里面，很多人就把它当做一个前端是吧？

不是不算编辑范畴，我这里因为实际上我们在做的过程中发现转换其实是非常难的，有可能比你后干的还是长，因为 Python太灵活了，所以我把前端的转换也放到了图片区里面，然后包括一些通用的优化，包括一些n相关的优化，这个是to b at里面做的，然后算是编译。

其实主要就是做一些循环优化，方程内存l它必须优化，这些东西最后一个就可以进。

当然大家也比较关心，就是说在编辑里面这种r的定义是吧？现在大家图片看到比如图r二那就很多种，包括最早的像t和lo，我们可以认为它是一个class的方案，就是一个非常朴素的图的定义，然后后来像他就是不认可做的是ssa的一种风格的，然后像我们的马斯报它是一个图的，还是主要就是为了更方便的都自动微分是吧？

然后探索里面用的比较多的海莱达，包括底下当然大家现在比较关心的是说我们有没有一种统一的r的基础设施加m二是吧？这个时候对整个AR编辑范畴的一个理解。最后我介绍一下就是说对AI边际未来的一个趋势和挑战的一些思考。我想就是说整个其实人家编辑的驱动力无非就是三块，就是一个应用的变化是吧？应用会怎么样的发展。第二个就是开发者的体验怎么提升。第三个就是硬件的变化，最后就是一个开放的问题是吧？要不要去讨论？

我觉得大概有4个大的课题，然后在应用这一块，其实上午大家都提了一点，就是大模型这块我就不再赘述了，但是这个里面大家的可能关注的点是不一样的，我关注的是说从框架来说，怎么样能够把这种模型的切分做到自动化，完全的自动化，对吧？

因为我原来是做数据库背景的，我做一个credit从来不会说是要用户这个切来切去的是吧？我这一块呢肯定通过call的model，然后拿到一个快的，plan就直接就执行了，但是我发现现在深度学习的做的大模型太痛苦了，是吧？这是一个，所以说我认为超大模型自动并行，不管它是能不能做到真正的这种病情，还是做到中间的某一个程度，这个是未来非常大的挑战。

第二个挑战呢就是说其实在深度学习里面CV跟LP的模型，其实那个结构我觉得相对还是比较固定的。虽然早上林总也讲了是吧，有那么多的变化，但是这个变化变来变去还不就是穿光或者CM那几种。但是未来我觉得很大的一个变化在于说，我们也要跟科学计算的一个结合，就说你在一个模型里面，除了做深度学习的模型以外，你可能还要做一些科学计算的一些复杂，这个是我们目前在做很多这个项目时候碰到的。在科学计算里面有什么特点？就是总结起来就三个，一个叫高危，第二个叫高阶，第三个叫不规则。

高危就是说我。数据的维度特别高，这里面一定要做大量的习俗，然后第三个就高，第二个高级，我们碰到了好多二阶微分的场景，现在已经我原来在深度学习里面遇到的比较少，二阶甚至还碰到了一个三阶的这种产品，其实对我们这个框架的挑战是非常大的。

第三个不规则就是说因为我们其实深度学习有一个非常好的抽象开始，这个东西你所有的操作都是探测产品探索和探测加探索就在范畴里搞，但是在hhpc里面，其实它不是这种范式，它其实是但是有index的时候，一旦有index，你这个规则我们现在所有的优化其实是不够的，我觉得未来对我们编辑非常大的三个挑战。

然后第二块就是开发者体验，其实我们现在的框架偏向于一种方式，叫动静统一的执行方式，其实实际上比较简单就是说我要静态图的时候就在加个艾特修斯是吧？

然后其他的那些动态图，其实这种方式我个人觉得他不一定是开发者最乐意接受的，因为没人知道，没有一个开发者会知道我去加了修饰，不由这段代码能不能加速，然后就究竟加速的怎么样，所以最好的方式你其实还是不需要加，就是否是吧？这是一个。

第二个就是说怎么样做到就是说我们许多芯片，其实它最好的执行方式就是把整个土壤到芯片里面去，我静态图什么不要动，这两块我觉得难点在于我觉得有三个趋势可以考虑，第一个就是我们能不能做一个跟动态图基本一致，比如说它的大量的语法，我都能够翻译到我们的静态图上，这里比如说比较难的比较控制的，我怎么解决是吧？控制的怎么样翻译到静态，然后动态立体怎么搞定，然后还有很多开始你的品牌的数据类型，这是一块。

然后第二个就是我我在动态图的模式上能不能做一个自动的线，就是说你不需要你去加这些思路的方式，当然现在有些探索可以，结果在底下回来自己来决定哪些热点的代码进行加速，但这个很难，因为目前看我们实验的结果就是编译的时间比他的执行时间还长是吧？所以说他因为自己很难系统很难知道哪些代码是能够做到自动的建设，但这是一个比较大的方向。

第三个方向其实大家上面也提到了，能不能做一个比Python更好的前端，但这是一个非常开放的问题，就是说可能需要比如说一个大厂是吧？需要想不想去搞这样的一个生态。

然后第三块就是硬件dsa这一块，其实总的来说就是跟我上午说的目标是一样的，我们能不能做一个图灵晚辈的，然后性能还可以的这样的一个编辑，就是说你提供一些原子的 plus的算子，然后用户可以在Python进行组合，然后我的编译器还能确保这些政府制定的扇子能够生成，而且性能还不错是吧？

如果这个目标能达到，我们就彻底的解放了开发者，但这个事情很难，这里面比如说像奥特斯盖里奥，特泰里这些事情都是非常难，而且你们可以看到整个看下来这里面一个非常通用的非常通用的一个挑战，就是cost model，就是说我无论是在自动的图切分，包括底下的饵丝Guidi，包括tywe里面，我都需要一个非常好的靠发展，这个也是我们非常难，而且我觉得现在也还看不到有解决方案的，就是非常好的解决方案的这个地方。

然后最后一个开放的话题就是说，这种MR的VR边际的就是实施的必要性是吧？我们现在已经其实我们公司也做了好多探索，确实也发展了，发现很多问题是吧？比如说mi的编译器就比普通的编译器卖编译时编译的时长要卖个10倍什么的，然后你还有其他的问题，所以这一块可能是一个开放可以讨论的地方，大概内容就这么多，你现在写好像我知道了好好谢谢。

说话人1 18:15  
有非常多的问题和思考，然后一会大家想要加入学工总这边讨论可以可以微调过去，然后我们下一位请梁军以前是参与深度参与的，管理派平台的研发有非常多的思考，也是我们群里的或者是大网红是不是？好，杨军老师欢迎。

说话人5 18:45  
然后谢谢杨老师邀请了，然后我 slut做的比较简单，化一页图，麻烦可能到下一页。

对，可能我这里边的话会说的稍微多一些，更多希望有一些互动，因为可能在dislike里边我不会太多很具体的一些工作内容，然后这个图的话因为刚才雪峰其实已经提了非常好的一个背景，包括编译，这里边其实我想我应该有一些不同的书给出来，可能是对大家可能会有一些启发性的。

其实整个从架构来说的话，我会其实大家都会把ABC分成大概我们前端终端、后端三个雷尔，不同layer的话，他们所面临的问题，他们的取舍也都是不一样的，所以我会沿着三个layer的话分别来阐述一下一些我的一些思考。

另外就是说可能除了这三个layer以外的话，当我们有机会从软件系统看系统设计的话，哪些放在硬件里面放在软件里面，我也会给出我的一些可能比较framework的一些思考，对，那么首先前端就前端，其实现在我觉得比较有代表性的东西，我们说国外的框架的话，那么可能看佛牌套的债算是目前比较hot的框架。

当然国内的话包括像once包括i don't know，包括像 MP，其实会有很多的一些我们会做的非常好的工作，包括技术，对，那么其实如果我们从它整个的整个的问题的原则性的规律的话，我会遇到包括几类比较大的问题，第一个就是说我的算子科技度的问题，从咖啡时代慢慢引化过来，到了更富时代的话，我们看到了一个问题，那么在咖啡车的话，我们可能可以非常方便的去磨改咖啡，我们加入一个雷尔，但是我发现每当我的算稍微复杂一点的话，我就在一个新的内容，这样的话其实对于整个我的model sharing，我的开发的效率不是很高，于是就从后开始的话，可能会是用以更细的算作为例子来描述模型，再进行这种定义型，做一个标题的模式了。

当然phone model其实它的模式不是原创的，其实在那之前的话 Cr已经做过类似的事情，就是没有像丰富推广那么多hot而已，对那么这个方式它带来表述灵活性的同时的话，也带来了相应的educat，比如说我们带来的这种客流浪子的开销以及仿真的开销，那么这个问题可能我自己个人判断思考是说在硬件加速器流行之前的话，还是可以接受的，因为我一次放松一次浪子开销跟我的整个加速器的盖没有那么庞大，那么它的灵活性我是可以去媒体编制的，但是随着GPU随着v100的分数call，随着现在的像a一版，包括明年NV的下一代硬件，包括现在很多di的出现的话，我们发现蒜子刻意度过细的话会带来很严重的问题，我可能大量时间花在了放存和等待上面去，那么这也相应的会有很多的一些软件硬件的手段，希望解决的问题，后面会稍微提一下下。

那么这是第一它校本颗粒度带来的一些优势和一些education。

那么第二个就是说我们算的覆盖度的问题，那么这个问题的话其实也挨过雪峰的观点就是图灵晚辈其实是我们对一个aabaa框架希望的要求，但是挑战就是说我对于新建来说，我怎么能够做到图灵网杯，举一个简单的例子，可能像它跟to是做语音比较popular的一个model，那么这个model比较有代表性，它除了典型的一些深度学习的算字以外的话，包括ft包括s那么这些东西都需要我们去提供支持，那么每当我更换了我们被骗子以后的话，我的硬件以后我都需要考虑是要把票所有这些东西，那么这种覆盖度的话，其实对于这种编译器的支持的话获得了蛮大的一个挑战，这是第二个。

第三它就是一根帽子，这个可能是版式的问题，可能是软件不兼容。一个bored本身的话，其实在chain这个时代的话就比较流行的一种模式，那么其实对于我们去所见即所得交的开发有效率，但这个模式其实在加速器越来越流行，硬件越来越powerful出来的话，其实非常有挑战的一个事情。

举个例子，前段时间跟一个朋友聊天，他说不能说是哪个公司的同事，他说句话，他说那些只会用一个梦的人，不会用100以上的卡，非常的激进的原因是因为在a版的话，我们打开搜索call以后，我的心里很powerful的，导致我大量的时间在交互里面去轮转，我其实受不了，那么包括像有些硬件公司的话，为了去解决这样的矛盾的话，会选择一些看起来不是最优点或者不是最优秀，但其实这个硬件来说比较重要的方案，比如说可能像某家新品公司的话，它的每次看到浪子的开发很大，因为它是要把它整个坑做一个在线的编译，despite知道很多的tile上面去，那么做这个工作的话，一次开销可能会是以秒级为单位，对于这样的场景的话，他恨不得说把整个组全换到我的家族里执行完，于是他就选择了一个比较看起来比较特别的路径，他可能是拍套式的话会把拍套式的枪给抠出来，自己做完所有的office，做完所有的mathematics说这个图对我全吃进来能够获得好的性能，但这样相应挑战就是说损失的我在算子的灵的灵活性覆盖度和通信这块的一些优势，导致我在换个场景的话，它销售大量是为了工作，那么这里面本身也是说一个model在带来我们前面表述灵活性同时带来的一些对于软件的挑战。

第四的话感觉是其实上午已经不止一位嘉宾老师都提过这个问题了，那么等于这个事业部其实是一个在这种AI现在的模型里边非常常见的一个场景，那么我一个推理的downside变化是比较辛苦的一个case，那么我这里开始models我一些帮的变化都会导致我的整个的计算能规则，那么做优化其实有一个特点，其实我们希望假设这个问题是固定的。

说话人6 24:22  
因为固定。

说话人5 24:23  
的问题的话，我们是能做很多次把什么类似的一些优化，但是当我的系统不断的变化的话，那么就会break很多的一些假设，所以当我在记得三四年前，我们看到FC的三个是不是假设的话，非常不理解就很奇怪，等到自己有机会开始进入到更多硬件的系统，会发现这是个非常nature的符合硬件架构师的一种思维。

我非常希望这个事情是固定的。

说话人6 24:45  
不能固定。

说话人5 24:45  
以后的话我能做极致的优化，但是软件的思维跟这个可能是不一样的，于是需要在中间找一些communic。

当然可能像m二出现以后的话提供了一个路径，说我们能够在表示层描述等于是一个行为，但是发到门头里，他没有解决说我们是个变化本身怎么在灵活性和性能获得好的渠道，那么这是另外一个问题。

第五他的问题是说咱们肯说follow上午其实林老师也提到的，我的模型跟它flexible，我可能根据我的不同样本，我之间的分支不一样，那么它本身在我的单个机器的执行会带来一个挑战，那么对于我多个机器的话，可能会带来隐私的这种struggle问题，导致我的计算量的分配不是那么规则，于是对我的整个策略的设计，我的优化会带来很多一些麻烦的地方，所以这就是在我来看的话，是整个前端可能带来的一些有错的现象和一些挑战，所以在这里面先做一个大概的 Summ。

然后第二块就是我们在终端看到的一些问题，因为前端是面向用户的，那么前端面向用户的话，通常需要转换成一个相对来讲更为干净的一个终端的表示，便于我们做一系列的变换和优化，那么这个里面我大概把它称之为是一个hardware的优化，当然画风可能不是那么的权威，是我自己按照我的理解划分的，他大概是说我们希望在这个里面做一些跟硬件不是那么相关的一些优化，但实际整个系统拆解我们不能做到完全那么的干净，所以在这个里边难免会融入一些面的信息，甚至某些底层的哈德威尔的信息的话，需要通过某些机制去透传到终端来来定义做一些事情。

那么这里边我大概可能列了4个4类，我觉得比较有意思的问题，第一类是在终端做策略探索，比如说我们希望去做自动运营的策略，那么关于这类工作的话，包括像t12的，包括像nice ball就做很多工作，其实像阿里这边之前阿里的话也做很多工作，像赵丹松是我的前同事，他们在xc里面的话围绕着b音色做了很多的工作，所以如果大家有兴趣的话可以去找男生去讨论，然后很需要你讨厌的问题。

那么除了并行策略以外的话，包括上午有一位同学问到了廖主任的问题，那么其实也是本质上叫做策略选择，那么只要是策略选择的话，都很可能会变成一个np2的问题。那么在np二的场景下，我们怎么能够获得一个相对好的性能跟探索速度的推到off。

第三个是混合精度，混合精度这个问题其实我想今天可能没有寒武纪同学在，如果有寒武纪同学可能会有感触一些，因为当我们会想通过硬件追求一些非常卡丁爱的极限系统的话，或者克服我的工艺限制的方面，发现，我就会带来对我的整个的策略更多的要求。

那么航母机的这种低精度的话，其实在我来看是这样一种探索，虽然可能有不同的一些观点靠的这个事情，但从技术本身来说的话，我认为是一个非常有意思的探索。

对那么在策略探索这块终端的一些事情，那么第二个可能会比较的耳熟能详一些等价值变换，可能像case这样的工作就是应该非常好的代表，包括你有些别的工作这块我就不去展开了，我觉得可能大家会更有些common的认识，对。

那么第三块就是关于children的问题，因为对于这种现代的新建的话，往往我们需要做很积极的failure来缓解kernel的态势以及访问开销的问题，那么我们该怎么样做非洲人？

所有人本身其实他确实是说像上午有一位同学提问问到的，包括是尹老师提到的不是越来越好，当我feel速度很大的话，我的整个慢慢的消耗，我的iPhone的消耗，我的GDP的消耗会导致我的b路的限制，可能反而倒闭说我虽然获得了我对他有意思，我虽然减少浪费开销，但是我觉得整个硬件的利用率反而下降了，所以其实这里面存在一个吹到我们要看到我们到底对不同的模型不同的此图，我们应该在 feel非洲plan空间里边做什么样的一些探索，那么这个就引申出来说一般来说为了做这个问题，我们需要有一个cost model来定义 cost function，帮助我们去做一些枚举。

一般来说在这一层的cost model的话，我们可能不需要做的非常的复杂，原因是因为这个空间比较庞大，所以我们希望能够做小块的一些，那么这块其实之前在阿里巴巴的话，我们同时有一些工作，可能大家如果有兴趣可以有一份大概是我们在x的终端的话，我们基于cost model做了一些非常大的探索，是一个启发式的类似DP的算法，那么这个终端可能我看到了一些有趣的问题，那么最后到了后端，这可能是跟哈尔威尔跟芯片相关的东西了，那么因为我不是硬件的专家，所以我就非常粗暴，我把硬件分成了gdpodiccpo这么几类，其实如果分得更清晰的话，我们可能还可以说什么c加a这样是一个特别的流派之类的，这些我通通就不管了，我就先把它简化成这个情况可能对于我描述问题，那么在WiFi这个层面的话，我们需要做非常多的跟哈德威尔相关的一些优化。

那么这里面我也列举了大概是5类我觉得比较有意思的问题，或者说我认为目前没有完全解决非常好的问题。

第一个就是说我们怎么样去描述张亮计算的空间对，因为在我来看的话黑暗的也好或者别的一些或TM的之前的版本也好，他们诞生的时代张曼计算没有那么hot，所以其实整个的抽象的话是以寄存器以这种非张量指令为例做调度优化的，那么这种调度优化的策略，当我们在张老师都会发现，它的整个的灵活性探索的便利性也不会非常的方便，所以其实目前我知道的或者我看到一些情况，有不止一家零件公司在做这种张亮相关的客户站的话，都会选择不使用TM的策略，也没有选择走向MR或者说是FC的策略，原因是因为确实在快速演化的过程中里面的抽象不是那么容易做，或者说目前的抽象的工具做的不够好，我需要我结合我的哈德格尔的特性来做相应的抽象。

那么另外一个问题就是说本身硬件在智能计算这块的话也存在讲究，那么我可能是不是有一些更近的 buff，我需要选什么size，导致可能本身这个策略空间的话也有些不一样，所以在这个地方的话，其实大家会发现会有一些好玩的一些要在这个里面做抽象探索的问题的发生。

那么第二块是我自己的观察，可能开源社区相对关注少的问题就是张亮跟向量的交互的问题。因为一个I的模型不是只有张计算，其实上午老师提到的我特别有感触，因为从算法层面来看的话，我们很多时间花在蟑螂以外的一些不浪费遍，但是当你不断优化你的账单计算以后的话，可能你的向量部分是那些非常flash部分的话，就是会平静的，那么他们之间怎么进行交互，怎么进行并行，怎么进行传输，怎么构成一些比较高效的生产者消费者的关系，其实这里边需要去考虑的问题，目前我看到的一些主流的工作可能还没有cover这个to，可能是我觉得后面也许会有意思，催生一些好玩的工作的一个地方。

那么第三个还是QQ站，因为一旦到了跟芯片结合的话，我们绕不过feel这个事情。那么QQ站的话，本质上其实也是一个策略探索的问题，那么我的不同的科我的拆解的策略，我去应付的策略，当我组合在一起的话，本质上就是个左爆炸的问题，那么对左爆炸的问题，我们该选择什么样的手段能够在1个比较低生存的1个他们8个人里边获得好的性能，其实在这个里面我们需要做的事情，那会接到第4点，我们一般来说为了开拓这样一个问题的话，我们会需要引入一个cost model，那么 cost model的话它会相对来说就要更精确一些，因为我们希望 cos猫的开票更多一点行为的话，可以指导我们去做比较好一些客站的策略探索，这块其实包括在GPU上面，包括在一些芯片上面，其实都会有些探索工作再去发生这些事情，对。

第5个就是compiler端k的动静的结合，对这块也是我自己的一个感触的，就是最优化的同学有一个特点，就是我希望假设说我是在一个s at的环境里面做优化，这个环境我是独占的，我可以用所有的资源去做优化，但是在真正的工业生产环境里面的话往往不是这个样子的，可能会有两个vca的共享一块tpo或者共享若干个CEO的coa，那么他们之间串儿很严重的。

这个时候的话我们其实并不太希望把所有的优化都在静态plateau到极致，而且我们希望留出一个8分和一个托词仁慈我们在run time里面来进行结合。

这是我觉得其实工作后来想去思考t的话很重要的问题危险，当然这个问题当我们已经有了成熟的硬件，已经有相对好的性能以后的话，哪怕通过更好的让他更快的结合，能获得2030%的提升，因为这种设计带来的cost可能导致说暂时不能成为嫡flow，这是另外一个问题。

但从整个太平河的角度来说的话，其实我个人会觉得说恐怕的终端他们的动力结合会是个非常有意思的一个问题对，所以我这边大概就是说从我看到的三个layer的话，来去大概说了几个问题，那么最左边是我画了两条线，就一个是往下的硬化的一个线，一个是软化的线，那么硬化就是说当我们具备了整个系统的全站设计能力以后，哪些功能我们觉得往下挪一挪，放在硬件层面会更好一些。

举一个比较simple的例子，如果说我是一个对推荐的模型有强需求刚需的一个团队，如果我把我的引发一定查表的操作部分下移到硬件里边，能不能获得更好的提升更好的性能？其实这个里边是存在着一些策略优化的探索空间在里头，那么软化是另外的一个反例。当我有机会把软件硬件全程考虑的话， Maybe某些在硬件里面都很复杂的事情，我在软件里面可以给讲。

那么举个例子，其实最近ina刚刚放开了proceed的打榜的结果，他们在拍talk的过程里面的话，应该有了进一步性的优化的一个提升，如果有同事有同学去看里面的细节，可能会注意到，它在PAD里面的话引入了扩大graph，那么在0.7的版本的话是在mkh里边，那么NV和plus在1.0里面在拍照这边加格拉夫，那么库拉格尔的核心是说我希望用尽可能轻的软件、的一些手段我去medicate，我在这种一个note里边扣到了你的开销，从而获得更好的一个全站性能。

在我来看的话，因为在这个地方其实他没有选择说我在硬件里面做更多的工作，希望去fix这个问题，我希望在软件c层面去解决问题，或者他所希望的一些投入跟收益的平衡点，这可能是我理解的另外一个例子。所以说整个从这三个层面的话，我可能会大概是有这样一些的问题，包括一些思考，所以我的分享基本是这些，好，谢谢大家。

说话人3 35:19  
好，我们下一位我们先看问题，好可以啊，我刚提 cost model，刚才金总说 cost model很困难。

说话人7 35:27  
我不知道是获取 cosplay的困难还是使用上的问题，或者是困难，或许过去对于我们做硬件的人给提供cost model，就是你想要什么样的cost model，我们做硬件的人提供 cost model不是很难，我没太理解这难度。

说话人8 35:43  
其实这个靠什么？我们也感觉情况就是，包括电脑可以先说一下，你说的可能是非常细节的东西是吧？你是CM的这种但是包括对我们来说没用，我们主要的东西我切的时候的是内存，比如说搬到哪里的时候，这种比较宏观意义上的。

说话人1 36:03  
所以这个是有概括的，我理解这方面的境界能够提供的。

说话人7 36:09  
开始，然后另外一个模式上来说。

说话人9 36:13  
分布式硬件。

说话人3 36:14  
当然有关他比如说我通信的开销计算的开销是吧？

包括这种分布式又是另外一回事，现在怎么我理解三个方法，第一个就是我。

说话人9 36:28  
大型的这种分布式就是我切到，第二个。

说话人3 36:31  
我假定我框架我需要把这个纸图下来，这是抠出一个字。

说话人10 36:38  
这个也是一个很困难。

说话人11 36:40  
我拿多少个东西给你出来。

说话人3 36:42  
然后第三个高层包子是说我在做是该流程跟底下算是是该流程a0的时候，要怎么来开立怎么来时代的，这就需要这三个靠什么？

其实说句实在的，我们在这种dsa架构。

说话人8 36:57  
基本上都不完整，或者说有些根本没有的话，我就全是规则，你去看我们所有的像。

说话人3 37:05  
融资优先这一块，就可能大量的其实就是可以是吧？我看到一个imi什么东西但是是，然后我碰到一个也就是我就变成一个这个牌子，然后。

说话人9 37:16  
甚至我们再到a0的时候可能也是一种。

说话人12 37:20  
嗯就是说。

说话人3 37:22  
之前看到一些是说反正他有各种策略跑，跑完再反过来，但是相当于是你用丫的或者是那种这种化学学生把它的反应或者记下来或者学校反正但。

说话人13 37:36  
我理解你说这个意思，对，反正我们其实也遇到比如说做一些自动的时候，你可能希望做一些 Ap的拆分，然后或者是一些比如说假如一个com是吧？

你可能希望把它放到比如说多张卡上公司这种朋友com这个模型的方式拆分的过程中，你是怎么样去拆分 com？

说话人8 37:58  
其实这里面它。

说话人13 38:00  
就相当于针对它的性能的一个预测，对吧？它就可以。

说话人1 38:06  
好行，我们下面第三个领导发言来自于医疗科技的袁进辉，袁总也非常要恭喜获得 Dsc2021年的seal were对不对？一堆做芯片的大厂，然后确实没有是吗？仿真 Top30对吧？好的好我们欢迎。

说话人2 38:28  
对。

说话人14 38:34  
对，也非常高兴有这个机会来做一些分享，然后这个也是准备就10来页，然后想讨论三个问题。第一个问题是杨军其实刚才问了软件和硬件这个是吧？就是说怎么分工，我们觉得也有一会我也会讨论一个例子，我们觉得有的事情你让软件做了芯片可能可能没有必要那么复杂。

还有一个问题就是说希望从我们的角度来说，来判断一下什么问题是已经解决了的，什么问题是还没解决。如果是已经解决了，就是说基本上就是可用的状态了，如果没解决，呃也许大家一起想办法怎么把它搞定。

第三个问题就是说也许再一个就是说也许框架就是说即使像这个框架发展了好几年了，像txt这种或者拍照的这种。

可能很多人都已经觉得写出这种代码的人都是高手，他做的都已经是无懈可击了，或者挑剔不出来毛病了。

但是我想讨论几个例子，说其实有可能再去回头去看的时候做的还蛮糙的，特别是当你有新的视角去看的时候，比如说我原来没考虑这种大模型的需求的时候，也许有些问题不那么明显，但是有那种大模型需求，你有一些新的这种视角进来的时候，就会发现原来很多东西可能都漏掉了，正好其实涉及到我前一段时间写了几篇博客，蒋老师就说你把 Summer is一下，然后正好来作为一个引子来讨论好。

那翻到下一页。我自己可以翻对吧对对对引子就是说现在大模型训练GDP3这种画了一下，英伟达他们做了一些实验，用麦克风来做的，你会看到用了几千张卡，300多台机器，分了64个组，每一组是一个流水线中的stage，所以它是分了64个流水线，这样的stage每个stage是6台机器，机6台机器之间是demo票了，数据变形，因为机器之间的带宽还是比机器内部的带宽稍微低一些对吧？

然后机器内部它是 Model怎么其实是做tensor的划分的， weight的划分，所以你可以看到它整个的所谓的神经网络像集群中去映射的时候，是一个蛮复杂的，可能换一个模型又换成另一种做法了，又回到说就像上午还有尹老师也提过 c DR a等等里面一些问题，我觉得大家只是在力度上不太一样，但是本质上是非常相似的，就是我这个逻辑的东西和物理的东西怎么映射，怎么排布，怎么路由等等这样一系列问题，所以这个是作为例子就引出了这个问题了。

好，这里还有个问题就是说刚才提到的比较复杂，是编程的复杂性，还有一个问题就是说运行效率的问题，运行效率问题同样的特别微观层面和宏观层面，我觉得也有很多的对应，比如说以前在搞芯片里面，大家都知道memory对吧？就是说我计算的非常快，但是仿存取数据取质量 lessons要高的10倍到100倍，怎么解决？以前有case乱去流水什么各种各样的编译器体系结构各个层面都做工作，在这个宏观层面其实也非常相似的一些问题，就是每个芯片它算的是非常快的，现在每个GPU上面有几千个扣，来一小片数据，它可能是几十毫秒100毫秒就完成了。

同时这些靠访问他自己的什么investment，带宽也非常高，就是p8的这种这种带宽芯片之间马上就降了一个数量级，你用n个link差不多是一两百g这个bug，然后服务器之间你如果用rdma的话又降了一个数量级，当然像刚才提到那个例子，英伟达它是一台节点之一台，服务器之间两台之间用了8根200gbps的这种高速的传输，所以它能做到集体之间也和艾米丽克的带宽差不多，但总体上就是说你会看到这里面也有这种hierarchy的现象，再加上深度学习，它本身有力度，大家知道是随机梯度下降是吧？

它和以前什么my reduce SPARC的处理任务就有点搞的不太一样，它这里面力度特别小，可以想象在这样的集群中，每一瞬间都有很多事情在发生，在这里面就有很挑战的问题，怎么去包括那些计算和传输，怎么让这些计算感觉不到所谓的饥渴对吧？

每算完一个，然后立刻就能开始下一个，我只是抛出来这样一个类比，就是会主要目的是说我们可能在宏观层面遇到一些问题的时候，也许可以在微观层面找到一些答案一些想法，所以从综合到说编程的复杂性还有效率的问题来说的话，希望实现其实对用户最友好的一种做法是说用户写程序的时候，基本上是逻辑层面的一些联网的多少层怎么连，具体逻辑的网络怎么到集群里面并行，怎么overlap怎么划分，怎么 Placement等等这些问题。

如果我们这个框架的编译器什么的这些都能搞定，就有点像你就打个比方就有点像硬件的一边一样对吧？在设计新建的时候，布局连线都可以预先做一些很多的工作，就可以一定承担解决的问题，这个其实就是需要通过运气，电器可能是既有宏观层面的，也有微观层面的来把逻辑的东西和资源的东西再生成一个所谓的物理的执行计划的时候，能够找到这种一方面是自动生成，另一方面是找到尽可能优的方案。

我总小结了一下还是这个例子，比如说像前面有数据运行，后面有模型运行的时候，实际上中间涉及到一套的，数据路由对吧？数据路由，当然我们是现在已经知道了是有规则的规律的，使用什么模式到另一个模式切换的时候，中间应该是插入什么运算，前项应该插入什么运算，反向应该插入什么数据路由这些，现在其实很多人都很清楚了。

所以总结来说就是说我觉得如果给定 Payment，就是每个oki到底发到哪个设备上去，比如它整个所有的设备可能分成若干组，每个AP也许发到所有的设备上去，也许只分到一个子集上去，这是places的，然后他发到这一组设备上去之后，它又怎么划分是吧？

它是切分了数据还是切分了wait等等，我们它叫Paris文化，自动的，如果给定前面这两个的话，怎么自动生成执行计划？

实际上我觉得这个问题是已经解决了，因为无论是从框架层面自动生成中间的路由方案也好，还是用手工的所谓的布线或者是手工的组网也好，现在都有例子了，自动的生成，我觉得现在跑的在比较靠前的其实是 flaw和mas bored是在kuang层面是这么做了，然后开漯河套时，现在还没有比较general的方案出来，然后态度flux在XL里面是有 tpo上的方案，但是GPU上现在还没有 worker的方式出来，这个事所以这个问题我觉得给定康菲的话， asthma和Paris的话怎么自动生成真正的执行计划这个问题可能都是worked dweller都okay了。

但是第三个就是说auto placement和auto，Paris仍然现在 paper里面我们都看到很多，但是真正的像比如说抛出来GDP3的时候，怎么能完全自动的搞定，其实这个还是很难的，但是又打一个比方，就是说在芯片设计领域，人家已经把这个问题都是变成work工作的，也许这个问题没有想象的那么困难，我们做了一些研究，初步的来说说，我们觉得不应该把它作为一个随机的一个概率的一一个很大的解释空间里面，就去不加任何的my knowledge去搜，这个空间实在是太大了。

但是当去做一些就理论的分析，比如说流水并行到底比数据运行或者模型运行带来了什么？如果去理论上来理解这个问题，就理解说流水平行它本质的问题到底是什么难难题，比如说它的核心其实是low balance的问题等等，做这样一些推导的时候，也许会发现比较好的解，它不是在这个解空间里面随机的分的它是，它应该是在整个节空间里面很小的一部分，那一小部分就是非常好的节，但是一旦离开这个很小的一部分或者叫机器学习里面那就错了的话，它就是性质就是解的效果就立刻下降了，所以也许在这里面大家可以共同探讨，我们觉得可以也许加入一些这种理论分析或者demand knowledge，也许用一些rule的话就能得到很不错的结果了，这是一个抛砖引玉的做法。

还有一个就是说单设备的自动代码生成，刚才大家也讨论很多了，这个也是现在来说是比较难的，也基本上是把它当成搜索的问题，对吧？不管是tdm去做，还是party去做，都是在解空间里面去解，只不过是一个是离散觉得是离散空间去解，一个是说用把它relax一下，搞成连续的再去解，然后再去找到一个整数的解等等。

所以迄今为止我觉得抛出几个观点，一个是说什么问题已经解决了，第1个问题解决了，第3个第4个现在来说有一部分可工作的解，这是一个。

还有说刚才提到的第一个问题，什么应该是软件做，什么是硬件做，我们觉得理解是从通过巴夫洛的时间来说，我们觉得在这种芯片协同的做这种工作就应该软件去做，不要用硬件去做，所以在上午他们还提到 Sarah那种超大的芯片，芯片上有很多带对吧？

同时我要编程的时候就要解决这些，这个上面64个大概这之间的这种路由又要解决内部的路由，就是本质上就是在解决分布式的问题，但这个比如说我们在框架层次实际上是可以减的比较好的，如果减的比较好了，其实没必要把一个芯片搞得那么大，你完全可以有很多这种功能非常聚焦的这种单个的芯片你可以做得非常好，然后同时他们中间的路由和协调就在软件层面就能就能解的非常好，所以我对 SQL内容就是以为例来说软硬件的分工，软件如果能做得非常好的话，特别是这种宏观层次的芯片层面，其实有些问题没必要在芯片层面来解决，而且涉及到分工的话。

说话人1 49:45  
就是像这种宏观的。

说话人14 49:47  
并行分布式，还有微观的单设备的代码生成，大家也可以分工每个层面。把每个层面的子问题解决的非常好了，其实最后一个整体的解决方案就很不错了，后面后面我再提第三个问题就是说，也许现在的框架看上去已经非常蓬勃了，但是实际上中间还是有蛮多 call的case没有考虑到的，特别是有分布式这种大规模分布式的时候，举一个talk的例子，这个例子其实在time flow里面有时候会发生，这个例子简单就是说有两个opo1和o2，他们分别需要三份检测和两份检测，但同时有数据搬运操作给他服务对吧？

数据搬运操作也是需要需要消耗显存的，调度的时候一般现在的框架的问题是在哪里呢？一个op是不是满足发射的条件，只看输入是不是available，对吧？其实现在所有框架都是这么做的，我看两个收入一旦 ireader了，我立刻就把包批发出去了。

但是发射出去之后，现在在一般的computer，比如说tallow里，在computer里面再去my output的空间，申请的时候预定好的话，当然大多数是能申请到的，但是如果申请不到的话，其实这个科诺实际上意味着发射出去，但是他看到了但是具体到什么时候可以继续往下走是不确定的。

这个时候这个例子其实举行了举了一个能跑下去的例子，但是还有的跑不下去的，比如说到了第二步之后， o二先执行的话，它一定是能先下去的，因为偶尔只需要两个，偶尔成功之前之后又释放出来更多显存，然后o一就可以执行了，但是o1和o2在2的时候实际上都是ready的状态，如果这个调度器先掉了，oe的话，里面就会出现一些不可预测的情况，有的时候 oe就在当前的schedule的县城上在做，就是依赖那种情况，他就把这个县城阻塞住了，阻塞住之后意味着也许还有别的ready的，他包括偶尔他也没法调度了，如果做得好的话是 oe它到另一个线程去标注了，它当然可以继续执行下去，但是意味着说另外到底要取多少个线程，其实后面还有一个比较复杂的问题，这这是一个这种dynamic，就是说其实在运行之前没有做这种类似于死锁的分析或者什么的，它整个任务能不能跑下去是一定程度上是靠急性外用运气的。

如果提前做一些分析，做一些静态的分析，整个流水线的发出来的分析等等，是有可能把这个问题解决掉的，比如说o一o二之间加一个可以dependency或者什么的就可以ok了。

还有一些问题说，还有一个问题，比如说举例子，就是说计算和搬运的，其实他们没有在现在的框架里面没有统一的表达，因为计算大家一般都是说有个计算图是吧？

但是搬运的时候实际上没有在搬运的操作，没有放在图里面的，但是这个是telephone的paper，大家会看到他在搬运的时候会插入三和receive，但实际上只是在逻辑层面把插入的三个CC，但是背后比如说跨了机器跨了设备的时候，背后底下真正的cw daman copy或者什么的，那些是不是三个receive来完成的，是在三个receive下面又触发了一系列的call back回掉， call back就不在于它 graph或者graph sq的管理范围之内，这里面其实会有一些比较tricky的问题，比如说刚才我们提到了为了解决死锁要额外加入dependency对吧？

这时候加入一些expensive的时候，demand的时候就会使得一些ot它也许有多重的依赖。

其实第一个问题就是说回调函数现在来说，比如说在框架里面的话，它只是解决一一个依赖的问题，比如我这个函数执行到这儿的时候就执行一个copy，但是如果我下面的运算有两个依赖，你前面的一个条件只解决了一个地方的问题，他另一个地方的没解决，这个时候实际上需要再到 graph schedule那里去查询一些另外的依赖等等这样一些问题，也会使得整个问题会变得比较复杂，这是把这个问题先描述出来。

一种比如说在探索里面有可能会解决这个问题，有一种可能的情况就是说 call back里面还要去查询 scale的各种技术器CTRL的状态，他也许会返回说你可以做了，当然就可以靠bug顺利执行下来，如果together告诉他你不能做，实际上意味着说 copy呢需要再包一次，再挂到 schedule的某一个微信里面去等等。

当然这是一个fix打个PAD，但实际上在现在的框架里面还没有连这么做都没这么做，也就是说他就没考虑所谓的中间的那些数据搬运的，假如有多重的依赖，假如我想控制它的顺序等等这样一系列的问题，还有一些就说这个时间我不知道过了没。

第三个问题就是说动态确定线条值，其实在很多分数系统里面都有这个问题，大家会经常用现场车来解决问题，但是 member put size一般都是here is确定的，但这个的时候也会出现一些问题，比如说我 size确定死了之后，也会可能会触发死锁，因为我这个任务的并行力度可能超过了你 marry threat铺 spread的个数也可能会思索，我当时前一段时间写了个博客来分析这个问题，还有一些情况就是说，如果说我 thread Poole是可以动态调整，但是他可能你不知道调整到多少，这个图比较抽象，但大概的意思是说现在的框架的scheme，其实最后大部分都是使用的最后一个，因为第一个就是说它在一个县城里面管理 graph的state，counter以及launch这些kernel，但是它有可能就会一个现场管理很多设备的时候，有一些context的，就是每个device要setter device，context基本上大家不采用办法，一般来说会让IRS每个device都有一个单独的线程去调度它，这个时候又会引入一些问题，这个细节我可能就跳过，大概我觉得经过你各种讨论之后，最后就会到最后一个状态。

就像 paiseh这种例子一样，每个device它有专门的现场为他服务，而且现场既管理 scheme的counter，也管理一些管理这些lunch这些科目这些操作，但是在这个时候，这里面核心的就是要确定 spread size，确定的核心难题在于说现在的框架没有保证，说我这个坑的浪池出去是不会阻塞的，如果没有这个保证的话，你就不知道到底要申请多少个县城是，safe是安全的，但是如果我们能够在框架层面去确保说我每个科诺组发射出去之后，它就是所谓的调度中的不可剥夺的对吧？

Non pre empty这种它一定可以从头到尾成功执行下来的时候，现成值的问题它县城的数量就等价于device，数量基本上等到这个就意味着说实际上就引入一个新的问题，就说我 wrong time的调度力度，不能是以前那种 stress的那种力，可能存在的就是操作系统EOS的力度，它需要是类似于高浪里面 Gorr rooting一样，就是useless Redd这种，你每个AP都可以是一个user thread，这个时候你无论多少op其实 cas thread大小是差不多的，对吧？

我大概是拍一些砖，就是说觉得说可能我们很多人会觉得说已经无懈可击了，非常完美的东西了，但实际上还是有很多细想的话还是有很多问题的。

说话人1 58:06  
好，我的分享就到这了，谢谢。跟总结在进入这个小组讨论之前，应该是还有每一位与会嘉宾都有一次机会来表达自己的观点发言对吧？那么我们按来，然后他们早上点过了对吧？那就是生根。行，那就来。

说话人15 58:33  
有size的，我们应该之前提前给过了。

说话人13 58:39  
行对，其实刚才吃饭的时候我还跟蒋老师说，我说其实上午本来想问一个问题的，但是上午问的问题人太多，所以没来得及问。

现在我觉得可以稍微问一下。其实我觉得是这样，就是说上午我们讨论了很长时间，我觉得听到最多的一个词就是吹走，是吧？

做权衡这个事情，其实我第一次听到这个词是我在大学上本科的数据库课的时候，老师当时跟我说，他说做系统其实就是做权衡，其实后来我们做系统相关方向之后，确实越来越感觉到做系统真的就是在做一个全对，包括像刚才说的上午说的，比如说芯片和软件之间它的边界在哪里？

包括比如说，软件的话它各个不同的层次，分别它的包括比如说编辑器里面还有怎么设计，放到哪个层次，这都是一些可能我们都希望去找到一个比较好的平衡的点。但是其实我个人是觉得就是说好的一个点可能没那么容易找。比如说至少我觉得就是说在短期内，我们有可能会很难找到，在这种情况下的话，我们有没有有没有一些别的事情可以做？

比如说像Mir其实我觉得他给了一个比较好的思路就是说我去找平衡点我可能很难找，但是我是不是可以提供一个工具给给给，大家帮助大家去寻找平衡点。

比如现在比如说假设我们要去一个走个路是吧？现在往东走还是往西走，或者是东西中间的到底是哪个方向往下走，咱们不是很确定，但是现在我先做一辆车是吧？我做一辆车的话，我可以帮助你去更快的去做探索。虽然我不知道这到底要往哪个方向走，但是你拿到这个车里多开几遍是吧？反正就慢慢的可能就把这个方向给探索出来了。

我是觉得说 Mir是这样的一个工作，除了那以外，就是说我们有没有可能还有别的一些思路也可以做类似这样的工作，或者甚至比如说ml它有一些缺点，我们可以就是说去做一个比他更好的一个工作，这个的话是我的一个问题。

说话人1 01:01:06  
好非常好，请下一位原告方也是我们吃饭的时候就在讨论的话题，对。

说话人8 01:01:14  
然后我这边也主要是在做MR相关的事情，然后刚才提到就说mlr的编译实际上可能过长，然后我感觉这是一个必然的事情，因为它就是一个多层l往下走，它层数多，它可能变得时长，它可能是这么一个调查的过程。

但是我觉得他这里面其实主要考量的一个是语义的统一，他把各种层次的语义统一了之后，然后我们大家可以接入到这一个系统里面，然后共享是一个体系的过程中的优化，就是可以共享这个事情。

然后另一个它就是可扩展的这么一个特性，然后我们可以在这对各种不同的硬件然后连到里面来，然后各种不同的前端也可以连到里面来，然后这是它的一个比较重要的优势。

然后然后我是觉得 MR这种还是比较好的一个想法，说他有没有什么缺点，他有没有什么缺点。

说话人1 01:02:15  
缺点在水合同里面，所以我建议你一会儿可以到那边，你们可以first呢就准备一下这个事情可以。因为纪。

说话人16 01:02:22  
龙我有。

说话人7 01:02:24  
两张出来好好。

说话人16 01:02:25  
还有其他的事情确定的品牌。

说话人8 01:02:30  
对到13个月嗯。

说话人17 01:02:33  
对不对？

我就稍微总结一下，我们直接翻到下一页的，就是说我们其实我做了三四年的这个东西，那一个compiler，然后分享一点就现在的问题，我从我们的观点的问题以及我们的认为的答案是什么，然后我们先稍微讲一下我们 dnf大概我们分我是觉得这样分就是说前端语言就像抽象编译方法意见抽象。我们从这4个层分，那么这4个层就是说现在普遍的方法就是前端语言，我们其实大部分都在做Python白点，然后计算抽象就是我们以AP为中心的淡，然后编译方法，我们其实刚刚很多老师都在讲，就说其实有很多编辑是个非常大的搜索空间，我们有各种的方法需要去找到到底是怎么最优的对吧？

然后硬件抽象其实现在我觉得比较缺乏的一个统一的硬件抽象，就会导致现在的硬件适配是非常困难的，那么产生潜在问题对于上层。现在的这种斑点的情况，就导致上层的语言那边的边优化和 a边优化一些分离，然后产生很多危害的 ok的力度，我们觉得 ok力度是为了描述算法用的，它不是不是适合去做硬件产品，所以这是一个错误的力度，然后会导致很多潜在问题，然后最大的超打算空间，那么我们优化有大量的编辑时间，然后最后就说是难以适合打印，直接到一下一页是我们最近呃，就是说总结出来应该所以一个iPad从我们观点觉得应该是什么样子的，分别上层的话，这样其实今天上午也讨论到了，我们认为就是说未来编译 DNA的编译应该是作为语言的特性就直接嵌入进去，然后但是不一定是走租赁这条路线，因为我记得谁提到一个生态问题对吧？

然后我就先不说了，然后时间有限，第二层我觉得更重要，下面三层就是说计算抽象，就是说到底 Dean的计算有什么特性对吧？就说他跟作为我们做df品牌的DNA和其他有什么区别？

我们认为是有迹象抽象，就像抽象，就是说我们是刚刚老师一直都在说说编译是一个比方搜索空间很大的问题，那么我们认为怎么去解决问题，从不是把它变成搜索问题，而是把它看成一个调度问题，就是说我们在一个形象例子要做原来做web news的时候，在一个非常大的class里面去调度的时候，为什么呢没有这么多问题了，然后我们现在也是想从这个角度考虑，就说我们事实上把一个 OPPO或者一个图调到gee上面也是这样，或者和对象有很多的基本单元，我现在想卖不上去，如果看到调度问题这个问题就会简单很多了。

然后这个工作是我们去年专门做的，其实是在做志愿抽象，我们试图把计算打成一堆小的task，然后去做调度，那么这个task怎么做还是很困难，所以我们现在在做的专门的第二版本，然后一个新的我们本来现在投 SARS还是有点不太热的，我们准备投今年的sdm然后稍微分享一下，就是说事实上我们需要一个下面的硬件抽象，我们现在认为就是说给所有的意见有一个统一的软件抽象，就它带有核心的可能有两个东西，一个是叫做我们一些胆量的一个探测不是问题就是说一些基本的就拍出来什么model，然后有了酱头软的抽象之后，就说相当于我们的所有的蒜子都可以有非常小的这种permit的组成之后，上面的一个好处就是说编译就变成一个白盒的编译了，就是说我可以通过主机会员组起来，就是把一个执行图变成一个SK普勒。

这样做完之后，比如说我的编译首先不需要去搜索了，比如说我可以做oneself的编译，然后其次就是说整个对上面的调度的空间也会化解很多，然后也就稍微回答一下上午每年的问题就是说我们试图这样做完之后，就说你上面的表达性不管表达什么样子的，下面可以直接就是一件的给你映射到GPU上面去，然后这样去做到了和 TM和现在的互联之间的一个比较好的渠道，所以我只是抛这么多观点，然后最后细节我们现在在讨论。

说话人11 01:06:44  
好，谢谢对待你。

说话人1 01:06:47  
到教育回头我们讨论的时候你意思好吧？

说话人18 01:06:56  
好，各位大家好，我叫赵杰，是咨询工程经济计算国家重点实验室的讲师，然后三我从三个角度来表达一下我对现在大家对应的一个一些理解。

首先第一个是2122或者是说语言对这种领域特定信息的这种传递能力，因为我们现在从模型到子图再到算子，在不断logo的过程当中，它这个领域信息可能会不断的丢失，但是你语音信息丢失之后，对于优化做起来，它可能就会越来越复杂，然后怎么把这个领域信息给传递下来，我想这个可能是 ma前期他想做的时候可能想综合进来的一个能力。

然后第二个的话我经常被问到的问题就是说，因为我自己本身做一些高科技的研究，然后有很多网友问我说这个topic到底有没有能够表达出一些t手工调度它不能够表示的这种变化能力？

我觉得这个问题它是这么一个角度，就是你自动实现的一个东西，它pm手工来去做的话，实现不了这个事情它是不可能的，只不过是你前期的这种 PR这种东西有没有支持，然后怎么去把无论是手工调度，还是说自动调度的能力去适用在未来领域，或者是说深度学习的算子future，他的这些东西上怎么把它组合好，是一个我们可能需要解决的问题。

然后第三个的话，我觉得我非常认可前面很多专家提到的问题，就是定义领域里边它要解决的是最重要的问题就是赤道的一个问题，怎么把赤道的问题解决好？是说白了我们在定义里面自己去解决的时候，自己设计一个 cost model，那么这个cost model到底怎么去设计？

现在我们自己想到的就是克尔莱斯克克里体，还有 recommendation之间的去到fps，然后现在包括上午我们老师提这个问题，就是说可能编译器它应该是以洛克利斯为主，我比较赞同我们现在最重要的一个优化可能就是说怎么把内存管理好，所以我们也就是也是一个方向，把包括Emme安排或者说我们自己做的 party的are做好自动的这种存储管理，这个是我们在做的事情。

我的观点就这些，好，谢谢。

说话人1 01:10:01  
赵杰老师也是今年硕果累累，刚刚发表了。

说话人19 01:10:07  
 pldi就完成了。

好行，然后下午的这几个专家的讲讲的内容，感觉基本上把深度学习编译的从前端到后端的各项内容都讲讲得很清楚了，然后我这边其实很难补充说更多的东西，但我在山东做深度学习框架来看的话，现在生殖学编译器要想在业务上真正用起来，其实很关键的第一步是要解决入口的问题，就是要能把网络或者说深度学习训练的整个过程能够用我们的不管是单一的are还是多层次lm表示起来翻译的过程，在以前可能按金老师说的以前可能都不在不在编译器里面，但现在来看很多原研究员写的那些完整的Python代码，其实我们都没有办法把它整个的放到一个电影体系里，我们只能做其中一部分，但我觉得我们不管是要做分布式的训练，还是要做底层的透镜，它都需要第一步你能把整个网络我们就整个训练过程都表示下来的，我就补充强调这一点。

说话人10 01:11:25  
好好。

说话人19 01:11:28  
我知道吧，你要不先自我介绍一下。

说话人10 01:11:31  
好大家好，我不知道，然后我现在是在未来的智能驾驶算法部门，然后大家看看我也是算做TM做的比较早的，应该算在178年开始做批准。

我现在还是很感慨，就是说今天上午听到了大家很多弹性的层次的东西，我之前在18年开始在阿里做的时候，那时候确实大家谈心是谈得很少，然后其实谈IBM的也谈的比较少，然后现在说大家我开始认同这个方向，我觉得听了很多人跟各个大佬也学了点东西，然后其实我今天上午我想想分享，就是说我其实听到像我这样讲了很多 dsa的一些东西，然后也谈到了就是说比如说像pm这种可能会有很多问题，然后其实我觉得说很多时候现在为什么大家在team上做一个芯片，就是说大家都还是能走，但确实会遇到很多问题，然后我觉得这个问题有时候不是说就是说为什么有这个问题，老师说我想分享是说其实很多时候有这个问题，是因为像我们前面社区是不知道你们这些问题，什么意思？

就是说每一家a芯片其实现在大概都是一种价格，就是说参数级别的邮报就是一个单位，然后waka一个单位可能还会加一个很简单的CEO，需要控制和scared，然后还会有一些自己的相当于一查因为去帮办事业或者回收的时候，他大概价格就是这样，但是他每一家确实有自己的独特一些东西。

在这种情况下，TM其实整个社区是很难去看住，就是说大家这些东西的就是说因为我不知道，所以说我没办法去抽象这些东西，所以比如说每一家可能在现在的题目上，他们想要做l去做的话，他们可能还会去对l进行一些扩充，比如说我会对我自己的这种每亩的这种层次去做扩充，这个时候很多时候如果是新的一家百姓公司，我也会做类似的事情，这种情况下其实会有很多重复工作，然后就会抱怨为什么基本上就是说没有吃，不是说天文没支持的时候，因为天文是不知道，所以说我这边其实想问一下，就是说还是很希望大家在做芯片的时候，如果有机会的话，能够将自己一些东西回到太原，然后我们可以一起来推动这个事情，因为我觉得像m这种他为什么现在能够做的比较好的一个原因，就是因为有发现这种各种CPU变更的，大家会运动，我们是去做，所以说大家去开发一个新的别的有一个蓝图去follow对吧？

然后就做的很轻松，然后我这次我讲的不一样。

说话人1 01:14:41  
好的，其实某种程度上也是代表 tmtm社区的一种声音，对吧？不是我干不好，是你们不会用你要告诉你。

说话人15 01:14:58  
好，你看呗。

说话人20 01:15:03  
大家好，我是来自百度飞翔团队的袁来祥，然后今天因为包括上午包括下午的环节，其实大家已经把问题讨论其实都很充分了，然后我可能看一下就是说从从其他一些角度，就是说编译器，当然我们提到很多的是一个性能优化，各种性能优化的方法，但是他边系作为整个一个中间环节，我认为它其实是有一个承上启下的作用，它前端是面向用户对接用户这一端，然后还有一个就是对接我们营业这一端，所以这两端的作用其实我觉得可能也是更重要一些，像我们最开始的可能流行框架是从静态图，基本上从静态图开始，然后大家可能有一批人也习惯了静态图片拍照是异军突起之后，大家可能都往动态图的方向开始转，包括TF和2.0的时候也开始变成动态图。

当然我们整个电信整个的架构呢，其实感觉又在往静态图的方向上又开始往回拉，然后但是这种刚才前面几个老师有提到，就是说可能很多动态性模式下的东西，它甚至是硬件不友好的，加速不友好这个事情，然后这个东西我们怎么把跟用户之间鸿沟给它抹平，因为用户肯定很只希望我写的很爽，我的想法和灵活的表达，然后我写完之后你又要跑得很快。然后至于你怎么做可能就不管，然后所以这块可能是我们也需要更多去考虑的一个问题，我们怎么去在支持用户更好表达的同时，然后把整个的优化给它做进去。然后另外一个就是说可能对接底层硬件这一端，因为我们团队也做一些硬件对接，然后可能我们做的就是说成本也会很高，然后我们跟一些硬件的团队去交流，其实他们做前端的一些设备，对软件的一些设备他们的成本也很高，然后刚才大家也提到就是说我们怎么去看分工界面这个事情上，就是说是不是能我们在编译器这一层上能达到一个更好的双方的更好的一个分工协作的方式，让大家的成本都降下来一些。

也比如说我们刚才提到的可能一些 cos model，比如说可能我们现在还是说两方面是有一些带的，这个get是不是可能从硬件角度他们也能更好的去解决，能提供给我们一些更好的信息，我们可能从这些方面可能能得到一些更好的结论，谢谢。

好的。

说话人10 01:17:34  
好大家好，我叫陈小强，我是来自新区段。

说话人21 01:17:41  
我之前是拿过他们的就是。

说话人18 01:17:44  
金总要做的领导啊。

说话人21 01:17:48  
然后老领导新领导都在场是吧？今天新领导已经走了。

说话人22 01:17:55  
然后我是17年在华为开始做TM做最开始的项目叫的，最后中加了一个故事，现在叫TV比较，但是我不是说你，然后后来马斯洛开始的时候，我就回了马斯洛团队，然后就开始做这个 leakage。

说话人21 01:18:17  
然后到后来出来如何。

说话人22 01:18:19  
去年的时候就离开团队，到新的公司做支持新股的 MP o值。

其实大家讲了很多，然后其实我感觉都讲的挺好的，嗯那我就从另外的角度给大家提供一些参考。

第一个我觉得现在在这个领域有太多的工作，然后有太多的观点，就是框架的ay变异了，还有硬件的就是各种各样的论文，各种各样的东西，就是可能让大家看花我提供一个我的角度，无论在什么时候我们要回到问题的原本就是问题的原本是什么，我觉得在12年还是13年的pidio那篇文章已经定义的很好了，他把schedule定义成了计算、通信、存储三三角形的一个隧道，就是今天大家所有很多人都提到这个问题，这是从学术上的一个定义，从工业界上对这个问题的定义是什么？

其实我觉得我打一个比方，假如说我这有一瓶水，对吧？

然后可能我有好多呃10块，然后我的目标是说把石块扔到瓶子里，然后保证说我中间留的空气最小，因为我们在部署任务的时候，其实要保证说瓶颈要被其他不是非瓶颈的东西尽量藏在瓶颈内，这样的话你把时空压缩的最少，你的执行效率是最高的。

所以其实在工业界大家做的其实就是这个目的，就是无论你是代表也好，什么算那个图算子也好，还是你是循环也好，你还是标量指令也好，还是控制了就是viw或者什么哈德鲁普各种各样的车辆。我们面对的软件对象其实就是一个带，然后但是把它展开之后，其实有这样的很多表格形式，然后你可以选择说这些抽象究竟是在软件搞定还是在硬件搞定，其实归根结底他就是为了把 Ply甜的足够满，然后把硬件的 resource他让他的resource的使用率尽量的高，所以我觉得这是很重要的一个一个理解。

然后第二个就是说我们可能很多人都想知道说a或者a kuang将来会发展成什么样，我是自己是怎么思考这个问题。的啊我觉得就看我们现在你写个c程序，然后你把它扔到gcc或者ln它其实可以非常顺利的四化的给你跑出一个二进制，然后你敲一个a点out，他可以跑出一个东西出来，然后你的程序有任何问题，他都会给你报 v或者l但是你会发现我们今天任何包括哪怕你在GPU上，你都做不到这样很难，这非常难，但是我觉得他一定是将来有一天会达成这个目标的。

所以这个是从这个目标来看，然后从架构来看。

说话人21 01:21:53  
我也觉得就是说。

说话人22 01:21:55  
aiba是将来哪怕是他语言或者编辑内部的很多东西，可能在某些程度上会继承或者发扬传统编辑的一些特征。

传统编译实际上是我理解宽泛的东西应该包括两部分，一个静态的编译器，第二个那些wrong time就是open ptvb还有各种各样的library库，其实这些整个加在一起才是一个完整的 CPU上的一个x8u的生态。

所以对于异构硬件来说，将来它应该我觉得也是类似的。所以说你看今天我们AI框架里有很多问题，就是他有可能他的思想解决这些问题的思想可能来自于传统编译或者来自于传统的CPU里的并计算对吧？我怎么样去做，所以poor的调度对吧？

我怎么用RCA的思想来演就是做异步的执行这些东西，其实从那些东西里面，我们应该可以汲取很多的营养，然后再说的比较细一点，我觉得现在有一个很大的趋势说，包括微软他们的 Wrong工作，我也关注他其实是一个我们发现其实现在算子和图其实在传统的框架里它是分成两级的，但是我们越来越发现他需要把这二者结合起来，就是说你需要有一个算子编辑来灵活的支持图层的产生的 schedule，然后这样才能方便你去把石块切碎对吧？

因为有可能你的k需要石块，有时候大有时候小，但你手写很难，非常便利的或者是敏捷的满足这样的需求，所以我们需要特色的编制来和图层的边区进行配合。

所以其实你放到传统的编辑里，你去看传统的从编辑传统的编辑器里面，它其实分过程间优化和过程内优化过程间，就是说我是一个放心call对吧？

我可以做一个net，然后过程内我就是指定我可以做一些指令调度，我可以做很多传统的经典的 csd其他就是产量这里很多优化这种过程间过程内，所以我觉得ABA器它应该也是会朝着这个方向去发展，当然它会有一些新的技术，比如说amai就是otg这些东西他会进来，但一定是说他会继承那些东西，就是传统的编译里边的一些思想。但它的表现形式可能会不太一样。

然后这是第一大点我想讲的。

然后第二大点我想回应一下 mi和TM这个东西，在我看来其实最关键的是说你要解决什么问题，而不是说你选择哪个东西它是重要的。 m二好Tom也好，如果你具备完全自己掌控，你想怎么改就怎么改，对吧？假如说你有像吴钊或者思源这种对7万或者天奇这样对奇怪非常深入的掌控能力的时候，我觉得就无招是有招对吧？就这意思。

说话人23 01:25:46  
就是说我们。

说话人22 01:25:50  
我应该专注在问题上，而不是说我是采取什么样的形式。

当然我也想给大家报告一下，就是说我们团队在使用器官的一些情况，其实我们在华为的团队以及现在我的团队，我们对TM都是深度定制，我们可以我从另外这样一个角度跟大家来讲这样的事情，我大概是12年左右开始在华为开始做编译器，当时是做一个在gcc上做一个DSP编译器，其实无论是gcc和lo m他们其实是CPU友好，就是他对DSP是不友好的，因为DSP里边有很多奇奇怪怪的东西，一是CD，第二是lw可能还有什么diss los，还有hadoop，对吧？

其实有各种奇奇怪怪的东西，你你你今天你去看lm，你会发现 Lm里面做软流水，他可能做的还不是很好。其实今天对于Tom来说或者mi r来说。

说话人4 01:27:04  
他们存在类似的问题。

说话人22 01:27:06  
当年我们在华为做gcc的时候，我记得我当时在骨头上搜的最多的一个话题就是how to put就是 to gsg architect。所以我当时今天我想肯定很多人也很关心这个问题，怎么样破你的npo到 TV或者是MR，其实这个角度是一样的，最后我们做出来啥样，我可以给大家报告一下，当时我们在华为做的。

说话人23 01:27:36  
您注意。

说话人22 01:27:36  
时间好， ok我就跳过，我就简单说几点，就说我们当时可能5个人的一个核心团队大概花了半年的时间，把应该是3个月的时间把o零跑通，大概是9个月的时间把o二跑通，然后一年的时间整个把技术ready，然后第二年又花了大概有15个人的一个团队来做产品交付，花了两年的时间把hi DSP，当时的6183现在应该是到6187，二十八八的我不清楚，但是那个团队现在一直都在大概这样的一个规模，我们把DS gcc给深度定制，我们大概因为gcc它是一个很完整的架构，我们他缺一些东西，但是我们加的pass不多，大概三四个的样子，然后在后端因为它有非常完善的后端，它给你暴露了非常多的call back就是who然后你去填这个户，就跟你架构相关的东西，最后ok你差不多就是这样一个投入水平，就可以把 GC玩的已经很丢了，机关我们什么我新到这个公司大概是花了9个月的时间，我们跑通第一个recit50大概是花了3个月的时间，然后9个月的时间我们跑通了 rest50和bird，同时它的性能可以跟手写模型非常接近，博尔特是超越首先模模型有5个点，然后波尔瑞士50的话是手写模型进入了95%，所以说关键说这个东西你怎么用，你能把它用好，就行啊无论是Mir还是TM他就能过了，他能给你带来价值就行。

我觉得我还想表达的一个东西就是说，其实我们今天大家讲了很多东西，其实有很多东西是在CPU和GPU上，但是在NGO上就是无论你用 Tom还是m AR，其实都是一个新的问题，只能说在华为在我们这个公司，其实我们对TM4npo这个事儿已经做的我认为是比较领先的，所以我们肯定是有很多经验，但是其他人如果没有涉猎这个方面，我觉得对他们来说是新的问题。

说话人1 01:30:06  
好好谢谢，本来是开源社区的布道者对吧？分享过去的开发经理，你非常深有感触，我都能听得出来是吧？为了论证一句话，你讲了从12年开始讲的挺好的。

说话人14 01:30:24  
非常有感触。

说话人15 01:30:25  
我就简单说两句，因为早上我也说过了，然后早上我第一个提出来有两个问题，感觉专一点玩，本来想抛砖引玉，结果就是run人人偏，了感觉，好像不是特别编译对吧？所以我就提两个我觉得我遇到的在做编辑工作当中的感悟，第一个是我觉得边开发也是有开发者社区的，为什么讲这么一个问题？

是我们最近因为第二项我们想做一个编辑的工作，我就跟我学生说。你去看看 to script，看看Emme，你选一个，你觉得哪个喜欢你就做工作。

说话人9 01:30:59  
我本来以为他会选择。

说话人15 01:31:00  
结果他本来我说passport为什么好，如果说一方面他是很方便，我很方便拿到他设计的rap，第二他是ssa我懂了蛮方便的，很理解就是说那边我看了半天还有好多层有点有点scary，而且这个学生也并不是一个说不懂system的学生，他是谁，他是原文的作者，徐主任老师认识就是谢志强。

对所以它是一个做system，student可以看到其实开发者社区这个方面是非常重要的，也就是说它不仅仅说我框架说的多好，对吧？还要营造社区，我觉得大家如果有时间的话，可以讨论一下说的是社区我们应该怎么去营造，这是第一个问题。

第二个问题是说我看了那么多就是编译框架，我其实一直有个核心的问题，就是为什么要有算字的概念？我觉得很奇怪，因为传统的编译器里面只有s也就是说我就只能去非常固定的一个大厂商，硬件厂商说我提供指令集，我可能要很多人觉得我才会很少去修改智能机。

但是在编辑编辑里面有svo这个概念非常奇怪，比如说我跑到msr，我去看我当时就关注msr，有没有定义好一个，比如说一个sat一个 study的site，说我可以根据这个来做编译。

说话人24 01:32:11  
发现它还是有operator的概念，然后我就觉得特别伤心，不过我当时看到清华的 G图，其实我觉得让我觉得工作蛮眼前一亮的，因为它有一个原算子这么一个概念，其实我当时觉得 form思路往下走可能可以可以做出一些东西来。

说话人15 01:32:26  
这个是我的一些思考，就提这两点，谢谢。

说话人24 01:32:30  
最近英特尔有一个叫cancer processing工作，也是就在回答这个问题。

好的。行，然后我这边是因为是做MAC，所以我从做软件的角度，相当于我不去假设进一步考虑做硬件的视角的话，然后提供一些看法，第一个是说对我m二r出来了这么一段时间了，其实我相信也有在场很多人都拿它做了一些实验，然后我目前对他有这么一个感触就是说他可能很多人会觉得他是个锤子，然后都可以砸一砸，但是他好像在任何一个地方没有表现出惊人的婆婆妹子，以至于让你很难受，我是用他的还是不用他的，但我想表达的观点是说或许是一件好事，就是说比如我们团队用做的两件事情，这件事是比较传统的，就是说我拿它作为一个技术的获得编译器非常简单。

另外一个是说我们整个微软件的CAD的interface拍摄的interface是用了ml的k波阵学去震出来的，就相当于比如说我在PAD里的一个Kwang，实际上那些kod是从我m二的一个一开始一点一点挣出来了，这个是它很方便的帮助我们就把从里面carno到cidi到Python，甚至到ps，所以这些层就都串起来了。

啊这是一个我觉得很有意思的点，所以我觉得它是一个锤子，到处都可以砸，它可能是一个好事，就是说它确实达到了它的目标，就是大家用它作为一个工具，很容易做一些相对容易的做一些奇奇怪怪的事情，也许谁砸的那一下就比较合适了，对吧？

所以我觉得这个还是值得继续尝试的，虽然大家可能都还属于官网状态，没有特别是做第二个是说上午也提到了传统CV算子，还有包括一些后处理前处理这些东西，现在其实真正工业界落地的模型里，我相信应该基本都是30%以上的时间都废在了抢货助理这些地方去，然后这些地方我们自己做的一些尝试。

说话人25 01:34:44  
然后发现。

说话人24 01:34:46  
这事不是很好做，不管是基于MR还是bm我们都做做很多的实验都 work啊，然后我觉得这个地方其实是如果谁有兴趣可以跟我们这边来探讨一下，这一点我觉得其实想象中是一个蛮容易达到怎么说产出的一个地方，但是知道为什么这个地方好像工作比较少。

说话人25 01:35:10  
对。

说话人24 01:35:10  
然后第三个刚才张老师提到的 coarse mode，这件事其实是一个很强的需求。我们平时在做各种优化工作的时候，其实对于一个算式到底怎么跑更快等等的方式，基本上绝大多数情况下跑这边试试，这个带来了很大的就是说，比如说优化的时候，我到底以哪种的药他要做，我很多时候我基本上要经过一个漫长的搜索时间才能得出这个判断，而这件事是一个我觉得挺疼的事情，所以想探讨就是不知道有没有人有兴趣说我们看看怎么能把cosplay这件事稍微往外抠一抠，不要每个人自己重复的做一套 cos，要么 cos特别还来，我要么特别多来我接不起来，有没有可能我们把它做成一个相对独立的，比如说在mm上面，我也一部分l我就能把 cos算清楚，也许是有意义的工作，有一些像清华sample他们有一些用来算 cos model to interface。

说话人4 01:36:19  
我这样刚才正好金辉提那个嗯提到就说 placement rock这块在芯片这个领域解决得很好，我想讲一下这个背景就是说讲这个故事在背后实际上可能会对我们我下面想表达的观点有些帮助，一个铺垫。

芯片领域 placement和rock这两个任务它是能够简单work，你要说解的好谈不上好，他是对问题进行了简化，首先就是说从芯片我们做a这个的话，我们整个发展曾经有一个很标志性的事件，就是说。

说话人25 01:37:14  
我们。

说话人4 01:37:15  
对问题对硬件芯片上的问题进行了一个简化，就引入标准单元库，标准单元库就是说整个一行整颗芯片还是有很多小的单元，然后每个单元的高度是一样的，因为我们嗯也这一个妥协，整个嗯把相应的软件社区给解放出来，就说他有能力来解决问题。

以前我们曾经遇到过一个公司，他想做什么就是三维芯片，还不是我们现在普通见到的这种三维，它是真的是说三维fpj然后说但是我要在三维的空间里边去找做place的主要体，这个东西对软件的要求就特别高，软件实际上针对一个特定的计算芯片来说，我是可以加手工干预，我能够找出一个解。

但是而且是说我任何一个用户，我有可能要涉及功能的芯片，有可能涉及功能芯片，但是你的工具都要能够给我解好，这个是很难的。

所以那个公司后来是fail掉，没有做下去。我想表达的意思就是说我们其实在讲编译，讲框架，然后讲芯片，所有这些东西我们来讲的时候，每一个环节的工程师都说我能够做到最好，做芯片的时候，我这个性能能做最高，做编辑的时候，你给我一个特定芯片，我能够把总能够给你编一个很好的 better那个框架来说，我自动分布式我做的最好，但实际上就是说每一个环节都有局限性，我们实际上是需要妥协的，所以从我的角度来讲，我是感觉说我们现在各个环节都有涉及，都在各个环节做过具体工作的人我们还是太少了。

其实我们可以说鼓励一些人做便利的人，真的是花点时间嗯跟跟硬件团队坐在一起，实实在在的分一块小任务，你去设计一下硬件，你去写一下，rtl一写完了你就知道，你以后不会再跟硬件团队的同学二个的时候，你说你那个不对你就得怎么样，因为你深刻体会硬件的困难，所以你不会说提一些不合理的要求。

今天同学也是一样，他可以说我去做一下软件编辑器的工作，之前，我在阿里做过一年一年多硬件设计，其实我的职业生涯主要是做软件，那一年是没有办法没人做，所以把我拿去做硬件做芯片开发，我不是一个好的一家公司，但是我在那一年我是充分体会到，就是说你一旦做进去，你会对硬件设计这块有非常好的一些理解，这样的话你再回过头来，我们当时做设计的话，我可以讲点，故事不会太长，就是说我们当时是说做了一个CNN的加速器，做了CN加速器，我们就是说当时是针对于特定应用，我们是把卷积的所有层我们都分析一下，它的计算赛，分析完了以后，我们就做了一个事情就是说我们用了脉动之内，然后我们决定说脉动之内我们应该什么尺寸的脉动之内去做。

我们当时设计的尺寸是7层、16层32这样的三维的尺寸实际尺寸，做这个恰好是对我们当时的设计是比较好的，然后我们设计了各级8份，这个8分就是ps还是要用的，这样的话我们可能就是input feature map我们要用8分，然后克鲁就是说克鲁我们也要用，然后output也要用，这样的话我们其实就是整个人为去规划了一点一个尺寸，然后包括数据的加载调度，应该是谁加载一次反复用多次，然后再加载一次再反复用多次。

这个东西我们这个规划也就根据特定模型特点来做，最后我们其实切换到第二个任务，我们到第二个模型拿的一看傻眼了就做不了了，为什么？首先就psms，珍贵资源根本就我们这个架构压根就放不下去，放不下去的话就说都会爆掉。

这样的话我们其实这时候我们当时应该同学就是去找了找燃力师燃色的工程师去找各种合适的芯片，没有找到一款合适的，当时怎么认为就做不下去了。其实这时候我们做了一个事情，就是把算子进行了拆解，那算是进行拆解了，但是硬件工程师他是没有胜似的。这个是你看你软件公司，你看坐在一起，你自然会想到说做这样的拆解是吧？

我们后来一拆解。

说话人5 01:42:34  
这个问题就解决了。

说话人4 01:42:35  
减了包括有些问题可能对硬件公司来说，他不会去往那个角度去想，他不是说他能力不够，他的思维模式有局限性，他就很难像软件公司这样去思考的问题，所以我就觉得就是说

说话人2 01:42:50  
哪家公司到硬件团队去做。

说话人4 01:42:53  
可以硬件工程师带来一些启发。

我们后来硬件工程师，我当时写了一个图，编辑写完了以后，今年公司是在接新任务的时候，他们就拿我的图片就各种玩各种改，他们就觉得很多问题就解了，对吧？所以我就觉得就是说我们要跨界，任何一个工程师要跨界，这是我的第一个体会，如果不跨界我们要讲说软件的边界在哪里？硬件公司是跟软件公司对话的时候就鸡同鸭讲了，各说各话对吧？所以还是要互相了解，然后这是一。第二就是说第二个就是说跟第一个就不太关联了。

第二个就是说现在我是看到说我们做各种。加速器，或者说我们做编译做框架的这种自动分布式，我看到一些这一块它都是很有特点就是说我这个就是做传播类的自动变形，我来做转机的，其实我是觉得就是说这里边就涉及到一个抽象层的问题，我们很多人很多我们没有把问题说，如果说我们越是解具体问题的话，我们可能效率会解更高解决更好，但是他懂大化能力通过能力就会越差，所以我们这个时候我们可能是需要一个平衡，我们要决定抽象成像的。

如果我们盯着的就是麦克风和向量预算，安排的运算，我们就盯着这些东西，最基础的叫我们关注最基础的计算pattern，和我们关注一些具体的抽象的层，可能这个是不太一样的。

我的理解我的想法是说我个人的观点是说就要像当初发展的路径一样，就说我们要退一步退一步海阔天空，我们不要说我们真的要追求极致，说我们要把卷积做的好多好，我不如退一步我就做麻木，那卷机它涉及到相当于是做数据整理，其实就变成没预算了，我是不是可以做这样的一些东西，然后很多复杂都交给软件，硬件就是说我把一个比较通用的计算功能做好，这是我的另外一个观点，其实我们现在像我现在做的事情，比较关注通用性，就是说一般的解法像搜索推荐一个讲就叫千人千面，每个人拿到的推荐结果都是不一样的，其实我这里边我的一个愿景就是说我希望做到是千魔千面，你来一个模型，我们没有赢的说一个并行方案对所有这个模型都是最优的一个方案，我们一定是说进来了以后，我们对计算pattern去做分析，我们看说哪哪种拆解它会带来额外的通讯开销，哪种拆解带来额外的存储开销，每种拆解带来额外的计算开销，我每一次对具体算子的变化，我可能都会带来额外的这种开销，这样作为一个cos，我可以把问题就是说量化来解。

其实说回到刚才讲的就是说我们解像ed行业里边我们解 policeman的那个问题之前，之前的解法是说我们 placement我们是要看曼哈顿距离，大家一估，因为有标准单元户我走都是横平竖直的走线，这个时候我可以很容易估出他的 cost，但是我们绕体我们用的a型算法， a型算法实际上他经常跟 policeman的他会有一个问题， place好了以后，可能有些节点它会有很多的连线出去，这样会产生拥塞，这个东西我们也是没有办法，就是说理论上讲是可以是能够绕进一起解，是能够找到最后一点，但是我们还是也是做了退一步海阔天空的事情，我们是没什么的先解，然后饶品后解，我们没有把它揉成一个问题，很多问题我觉得一个妥协的过程。

对，我就说这些。

说话人1 01:47:34  
感谢蓝松勾起了我PHP回忆，我也是一天出生的，也是一堆算法和硬件融合在一起的。

好，我们最后请高叔叔来说杨青大家好。

说话人11 01:47:47  
我叫高阳，来自于商汤科技，然后我主要负责推理引擎的开发，然后还有基本的就是核心的算子库的这样的一个开发。对，然后其实我也是对在是因为我们要在写算，不管是写算字，还是就是说我们做代码生成的时候要做这种参数空间的搜索，也是面临着这样的一个搜索的空间比较大，搜索的速度比较慢的这样一个问题。

然后我的想法就是说也是比较关注硬件的 Cost model的这样一个问题，因为我们生成的代码它也一定是就是说跑到硬件上，它实际上是有这样的一个坚持的物理的基础在这里，所以它一定就是说它为什么跑出这样一个时间或者这样一个性能，它是有它的一个原因的，就是我们尽量可以通过硬件能暴露出这样的一些参数，或者是一些体系结构的细节，然后我们就可以去推测这样的一个时间或者说一些参数的设定。

举个简单的例子就是说，比如说我知道了一条指令的吞吐量，还有延迟的周期，或者是各级开始了这样的一个组成结构，或者是它的一个运运行方式的话，也包括它的一些数值的这种这种参数的话，我可以在比如说在构造分块的时候， timing的时候，我们就可以大致上就有一个范围，或者说我们就可以排除掉很多不相干的一些就是干扰项，对我觉得这样的信息可能还是需要做体系结构和芯片的这样的同行，还有做编辑的同行，共同来关注一下这样的一个相关的一个就是结合的领域，对，这就是我想提的一个问题的。

好。

说话人1 01:49:48  
我们好我们看看时间正好我们3:30，就所有的咱们发言都单独发言结束，我们现在就开始我们现在就开始分组讨论，然后会相对比较开手一些，然后我们三三位主持人杨静辉和杨军你们得。

说话人15 01:50:07  
应该辛苦了，到我们这儿去吃吧。

说话人1 01:50:09  
我是学习我们自己的课，那静辉在做主持，你做我就做一点。

好的，然后这样的话这边就6个人对，可能这边还需要再过来1个人，雪峰这边可能还带来一个人，然后其实刚才有不少，其实我听到了一些共识，就是说各层之间，包括其实兰松和吴赵还有小强都讲了要先去融合先对话对吧？

其实就像雪红说的先垂直整合，再去考虑分层的事，我好像该到大概这个意思，大家好像都有这么一个结论性的东西，而且还有不少专家其实自己带来的问题像敏捷说为什么要源自这个事，还真是问到了，我也没想明白。

所以我觉得大家带着自己的问题就可以hold住这个问题，然后讨论的时候也争取把自己的问题抛出来跟大家一起讨论。好吧？然后我们现在就开始就按照分组来，大家可以围坐在一起，三一边三个都在一起，好。做。

说话人2 01:51:21  
好对。

说话人1 01:51:24  
我们讨论时间是一小时是吧？我们先讨论一个小时。

说话人7 01:51:29  
前面一块他还有没有让我操作也行，这钱怎么分好，对刚才那个好。这些问题好，对刚才大家提到比较多，包括我也不同意是吧？

我现在做的事情网上的。

说话人9 01:52:02  
这种情况交，但是我们大家感觉说是主要实际上就是说我们聊三次，其实你看还有有几十个犯罪群众他还会在增长，但是你可以去看差异的，30 34个300 200多条，然后比较稳定，而且你就对这2个就这么来搞应该没问题。

什么概念？我们300多条指令他是很有限很好的一个数据，是不是？他几乎没有多少更好算出来，实际上我们现在的话可以抓紧加入，所有的是因为这个算是达成它是利润性的高效的，是编程，我觉得他不是为了高校接待，他算是个边界高校是公司性的，还有各种需要的我们的公司高校计算的东西，所以我们一定是个大胆这么来讲，我们现在看怎么样，然后哪些工作在一起，然后怎么样实现一个计算组合怎么办？

我们完全是在说，我觉得就是说确实有些是比较糟糕的一个东西，但是他自己也是一个好东西，好多东西就是必须编制人员。

说话人1 01:53:28  
我现在稍微打断一下，大家聊洗净之前讨论完小主持人小组的主持人要再上来，大概要在对10分每个人不超过10分钟再讲一下你们讨论的一些结论。

然后其实刚才三位这个这个组长其实讲的问题互相坚持不是特别正交，大家思路切入点不一样，可能所以可能还是大家你们三位听完对方的这个介绍之后，可能也要稍微整理一下，比方说有有的，怎么说好了这个东西我们没有提前整理，所以现在可能也需要大家回来，刚才那些问题，我们最后形成一个。

说话人15 01:54:10  
杨军还有对我们有个例子。

说话人7 01:54:13  
但是你们讲的有不同的角度，不同的分法，这个问题可能有的有相互power的对吧？

有的看看你们三个能不能把整理，我觉得还是很有价值的。

说话人1 01:54:24  
好，今天我估计不一定问题太多，我们今天也可以把问题先整理出来，然后我们的对周期性讨论，然后有结论的可以有结论，没有结论我们就主要整理问题是可以的。

然后我们最后把三个表目的在一起。

说话人7 01:54:45  
你说东西现在是对行行，那上面的话中午昨天。

说话人26 01:54:59  
对技术。

说话人2 01:55:05  
好或者要我找到那个政策，基本上进入两个不同的安排，参加有沟通的调整上面，所以说大家做的时候呢，咱们就是说这条道路说是我们以后都应该包括他的工作，两两岸还有一定要有这个概念好。

对可能还没有什么，就是说现在我。

说话人15 01:55:51  
一个我们想不想不行了，我肯定不想去怎么表达，就是这种机构来我们压力差了，最终他还以后可以。所以要回来的办法，这是非常重要的一个简便的加上美丽，或者说算一个人是研究员，他专用盖加速器，研究你们说那么现在我们需要回答的改革为什么？

然后就。

说话人7 01:56:16  
在做4这次确实是很积极的，为什么要靠这两个问题，一个是AP，其他的各种算法公司看，第二个下进行演绎的时候，这为他们有些我觉得可能主要是大部分有些一点过程感兴趣时间分析的情况下还行，然后可能思考我们公司它有个问题，它没有一个这个矛盾，我不知道说说大家所有的上市或者说其他一些国家共同讨论的一个语言，包括我们的我们有时候学习包括有些问题，但是正好是一个看好的事情，我们现在正在好好的主要做的我们希望算法的事情提出来变成一个工程化的事情，我们说是一个非常科学家的事情，对这是允许当时的其实因为还有一个主要的方法，他就可以存在几个这块一个栏目的，因为会造成这种方法其实很多人也喜欢，以前是最早就是资深的地方就是这样，我以前搞一个number上表达实际很好，对一一直我觉得他但是这种可能对公司工程是编程的业务，建筑工作才能做出来，然后然后到底下确实因为到底你像编辑那个例子，大家可能最希望的说看到自己到肯定好，因为我可以就这样，我不知道我们这个东西 CP走进自然，我觉得有看到的概念，我觉得他其实是重新解释一下，就他好个人感觉的地方，比如我觉得以前很高兴的计算理解意思也是类似的东西，但是比较他的因为他们对一个跟其他这些也好，其实它都要加一个当时还需要你给我，我个人觉得这些电影其实实际上就表示对这一段你是要做定型的，你是这一段你是要做而且量化的，但是刚才我们什么还有问题问题没有，确实还是不出来，这个东西是这里说想法这个。我到时候就发现主要是所有的，而且更重要的如果能做到这种思维方式，就搞这个可能我刚才也提到一些包括查危成藏干的那些事，确定的他自己希望到底知道什么，包括规则的增加。

说话人15 01:58:50  
没有这个感觉时间上面是吧？但实际上嗯我们做很多修复杂志或者一些形式，其实都应该我觉得到时候还有一个这种原装产品，不成啊可以说是2012年。

好谢谢。

说话人7 01:59:05  
所以我这里我们已经跟创新的工作差不多了，实际上算这个东西我不是说有还是没有的问题，还是说如果说我深度学习我不要这样子，因为不在这个工作，不在里面前导出的问题，这边的价格现在不断的变成一些可能，所以其实这种情况很多意见，但其实让我看了4大的生产企业这个，他也有大量的问题，他也有偏c的这种方式，然后你可以用这个上面的刚才计算你在他也有用处的，还是其实用我们的本身我也可以把这个东西写出来，但为什么没人用过来去写参考这个问题，就原因就是在于上海工程它是很好的情况，上海或者是他不懂很多的文化，他们很难有钱给汽车合同代码运用，他们自己去搞，你不可能这么去写深入学习，所以反过头来深度的学习的三个孩子是，40多块钱和一大堆的系统工程师，为了让方法给你写起来给他，我把这个事情算出来算了之后，我本来应该说他是很大的时候，当时说还是调用的时候大家放心没问题，那么依赖关系我也可以说话，我也可以给你一种很泛指的表达，我们两个是这样的，我觉得他表达我甚至像一开始传媒你都可以写一个这个文件，你读的文件我就不知道他算是的形成，其实是为了到一定是要做一个牺牲，然后现在想法功能简化，然后在这个过程中又延伸了一个问题是当我有单之后，就是因为现在深度学习我的意见，我身边也去了，如果npoepu等等，我是不是说我这份代码的CPU上写的一份到QQ上面写一个，我们不能写，我就有他是不是要写的关系。

所以我们要说我们部分一个意见，其实你这个东西还是很复杂，所以扇子也能带来这个价值就是说我不管你用什么样的硬件，但我的判断看不懂网络上，如果要往往是我着下来，我知道网络你们采购领导在一般情况没有为了骗钱的这次这样算系统，这两个也能画出这种水平均衡发展，所以这个东西是最核心的，然后再从这个计划其实就是咱们就觉得更加有价值的。

我上次我觉得至少能但是最早期的圆环的就是说百分号先教了，原来就是我们，所以我们要编译，我通过编辑方式，我知道他规矩大概是真的是吧？这种一定来说首先看到从能够表达很多的话，这样我就可以传媒是各种各样的这个表达，所以慢慢的就产生了这些经历的，就是说我觉得这两点就是这样，你刚才提到系统叫做这样东西，然后等到你在哪去了，我觉得这条件很重要的。

对完备性肯定就是说，但是因为很多人还要去看看咱们市场的时候，最主要考虑到量化的时候，实际上也都是把量化的，本来就在会议。

说话人26 02:02:43  
中。

说话人2 02:02:46  
都是发的我的学校的学生。对。这样系统一块的一些。

对关键说。

说话人7 02:03:04  
大的情况就是这意见也是规定是我说出来，其实最后他一起到有这个项目，你看把设计算到这里，但是生意然后他就是变化的时候，我们说什么最好玩，我觉得什么自己吸收。

我那的就是研究就本讲法，所以这点把这些我们可能可能就会更好的，所以现在他们但是也是他们的对，其实我觉得基本上就是一个最贵的的产品，就本质上的情况。

你跟他们现在咱们讨论一个跟他一下，反正从或者说学生也好，或者说行，他在宁波市场做了一个其实那个产品是吧？

对，其实因为我们自己之前把东西给我，比如像之前进行他的麦克是吧，这种这个也是一个综合的资金2/4。

说话人2 02:04:10  
这个事情就是。

说话人7 02:04:13  
确实可能就变成了难易程度上面看到一点比他发挥自己的方法管理，对，但是他只是建议的话，他牺牲的兴趣的出来，对当然我们。

说话人2 02:04:29  
这个的话也。

说话人15 02:04:32  
是可以的，我觉得这个事可能也是，但是也是怕对，所以我是说咱们这个工作为什么首先这个方案这样，就是说这个也是说这个方案我们要参与的话，便宜再稍微易购点可能性放在这个里头，那就一点点不能加对吧？

只是原来比较通过的，现在你应该可能怎么样到最多的，我之前的问题就是你走的金融的解决易购，你先把他的课题至少我们先找一个人说，然后再往后再大家都能够参与，我们这会看这样大家的效果怎么样，然后有些我觉得就像是三个，但是你就是解决的变成因为周期比较，所以我就需要先不全部把一个东西就行。

好了，大家都这边再看了一点开展，你不要需要我应该匹配的这种cost，我觉得这个国家如果反正我觉得我觉得还有方法非常好，对吧？我们可以给你流程。

说话人7 02:05:35  
你想就是说直线的这样或者是面积，然后这边公司或者是互相的对。

前面你刚做完了，其实我们把它这个东西，比如说我可能会聊打下顺序，有可能它这种是一些交流也是会比较高，它这边是效果比较好或者怎么样，并我是没有办法做到说就在这里就有几家做下去，我们是本来我们这个平台的。

说话人15 02:06:05  
对，就是说现在我们想现在的要走了这样的，你把他们先拿到什么？

交给一个其实有办法的事情，你要很讲认为，然后我们这个问题大家都给你介绍一下，有三个观点。

然后刚才所以他们就开始我们翻译出来带到2020年，当时非常到这个地方的事情困难，大家最后面自己因为他不一定出来，因为这个话题我们确实讲的大小的地方可以看到非常感觉，甚至有的时候大家而且还要做出来，你稍微给他想一些东西，国家一点点就是通过什么下去之后，然后而且你没法分析去讲这个项目这个是在上行，是其实要最后你不知道你是有一方面应该说现在做的到底这个情况没有下调，然后这个里面其实针对我们最后的情况之后，他给业务的市场管理的时候，只要有了那些推进了，老师怎么会有机会这个环节就会发现就跟当时调度的时候，就是你把任务都打成小的专块，无非的话一方面其实按照一个内容，而且我们现在不知道，所以其实而且你看到每一个做IP也好也很多，从规则实际上它就是全县都还不是很明确，他就你这个问题之前那个药可能调动在消费产品不一样的，我们可以协商的时候，我们可以很难考核一个统一的结果，我们总结了你们搞什么，你们这方面没有需要长期的，没有所有的我上来之后如果真的存在的方向，原来如果要存在实现的话，高层可能一些现在我们去解决，但是反正自己的信息不好说，如果你们有权利行业的我们知道这个是中等职业，对工作中间的领导品牌什么就是这些服务，我到现在我们可能到时候然后去找，要想一个消费成绩到比较好，而且你可能专家都不知道，这样的话就说要求好看，在可能大家非常稳定的水平，自己的内心想要去跟他说半天，这好其实你方便你说我们就说点，如果在训练时间上面，很多常委是这样对话的，接入可能基础上还是所以说可能我们这个集团的东西。

说话人27 02:08:34  
基本上一个。

说话人15 02:08:35  
真正技术做的产品的，就是说现在我们还是需要，其实你可以把这话讲，我可能这一块实际上现在所以我们感觉到了转到了这些没有其他的选择合作，感觉教育其实和你还可以，我觉得你自己的问题要存在的必要和问题解决问题，同时的话好把往剩下的服务室或者更多的长的机构，从这个城市就会，但逼迫他是不是导致你的问题，我觉得当时的不一定是土地完备，但是说至少这个地方能够做到的话是不够的刊物去，而且考虑的动程公司的动产就不太能接受，说他比公司算一下，就是说这个核心就是如果专利用了，下一步把这个价格出来，这样可以不干，可能是一个就可以，大家都会觉得有点想法有点接受，其实成本你们的算采购也是他们的，所以让你们虽然这么说回答的他们就可以。

说话人7 02:09:35  
他肯定在两种选择他这个公司他都出现了，没有查到了，就是说也是两个，我上次有可能像指令一样，因为收敛的一个最小的就像刚才提到的这一块，如果是指定的，结果硬件的形成和他这种转向不一样，所以他说我们的问题。

说话人27 02:09:56  
就像。

说话人7 02:09:56  
一开始大家都叫累是吧？

后来一些就是说我发现那样就不太多了。

说话人27 02:10:05  
我估计可能再过个两。

说话人28 02:10:06  
三年感觉算是这个东西太急了，有可能就变成了。

说话人7 02:10:10  
所以现在你要发挥。

说话人27 02:10:11  
他打破。

说话人7 02:10:12  
这个图算的边界，或者说算是要拆对吧？其实他的概念还有，然后或者说他有在赶紧去帮你帮助，是不是没有其他意义？其实我们现在已经在想要算出来怎么就这样算出来，按情况那就行了，什么问题？是因为它是强制性的原因还是连接来是吧？但是他如果就像真的那种理想的情况下，你就整个系统只要实现十几个几十个任何上面的计算全能，通过他听了他就是指令集那种感觉了，做硬件也方便了，做这个事情也不是不行。

这个我确实。

说话人28 02:10:56  
一一对应。

说话人7 02:10:57  
的问题是吧？你一的我当然好办啊，你现在的电流中间，你没法一一对应，这就是中间人直接在我们这边，像我们的话我们做因为我们也做一个便宜的事情，其实在我们没有严格的算计的概念。

对，比如我们可能从一个班子里面来一个紫图是吧？

但是我紫图最终可能还得生成一个可执行模块的话，一定是会很牛的，我会在这里可以把它叫做一个事情，其实他这个东西动态就是把安定，然后让他至少在这里面他给他们是多少最低的这个优势或者差，或者其实已经在把它变成对，所以他就没啥意义了，你说其实也是看应该其实就看就配到硬件能力上了对吧？

包括你拆分也是因为严嵩增加或者都和硬件更重要的限制是硬件的实力，而不是算边界，我觉得是吧？是这个概念。

说话人9 02:12:01  
我其实突然刚才他们提的是，我。

说话人17 02:12:04  
没有的，我觉得什么太极我们都不知道听说过，对，他觉得同行我把谁叫来了，对他们很好例子就没有一直遗留问题，他一上来就做，你看他好像就没有ok概念，他就直接把写完之后他他的目标都拿来了，对吧？

然后他们选择一样，这个时候所以。

说话人7 02:12:28  
国外面还有挺多交流，我想就相当于我们自己把自己的思路不太一样，然后他他的好处是反正他从一开始。

说话人15 02:12:37  
然后一开始他。

说话人7 02:12:40  
他也是爱的，所以中间有一段我可以去识别，然后其实还是帮你做一个您做什么您的联系这块的微信的这些人，从写公式到出大问题的话，你不用再写出大啊，他把很多这种关系是就开始直接把这个扣了是吧？

反正我觉得他的目标可能比现在我们的目标还是太大，所以就导致挺难，然后他更明确的话，我觉得相对来说单反一下是吧，反正我们现在其实我觉得实际上就说到刚才说了算了，刚才反正从我们的感觉说，我们其实没有特意的去强调算出这个东西，但是就是说在实际上真正做的时候，还需要有一个算在为什么就是说就像我们刚才说的，比如说我一个pass的model是吧？

说话人13 02:13:37  
我们可能会把它比如说啊不是分析一下，然后会比如分成很多个子图，这次图的话，其实你每个编译完之后，它最终其实就是一个反正是我怎么去，因为没有别的担心谁叫，所以我们也叫上司是吧？

但是因为你不能把一张图完全的像以前比如企业家程序一样变成一个可持续文件，如果是如果比如说我整个模型，我可以把整个模型都编译成一个这些文件，我就没有上次来，但是现在可能包括我们在过程中也发现就是说中间有一些你说一个边线，假定新店会有些断点的点关过来，比如说他。

说话人7 02:14:25  
有一些可能有些是因为性格问题，所以我为什么讲这个东西，大家需要一个东西把也是包起来，然后也找不到一个其实啥？

不能把它当成一个整体，对合适的名字就叫对其实他没有看到的是需要资源编译里面行业根本没有svo，这个概念，是建议 ssa的方式，然后上次应该是从更高层图无艾尔之类的再往上比如特色flow里面这种专餐里面定义算组的给什么，然后可能就是因为从那弄一下，因为比如说 ml它就是从这个图上面穿起来下来，所以他可能就是沿用了，但是其实MR里面 operation就是ssa的operation也没有算这个概念，基本上都是粗力度的慢慢变成细力度，是这个算是你的算法人员，他听这个上面一个一个解释算法上，我觉得你知道理解这个算法是有些可以比较清楚，对吧？

徐老师。

说话人29 02:15:44  
你们工作里面把一个蒜子切成很多一个小的，是不是也可以理解成一个蒜子可以切成很多的小蒜？小算子对。

说话人17 02:15:55  
核心就是说我们把它一个算子市场天然它就是分成小块，你像一个克扰在基本上面执行的时候，它最后就是很独针，只不过这层信息原来没有暴露出来，我们现在就想把它变成一个很显示的一个抽象给弄出来，全部弄成小块对，并且我们希望每个小块的资源就你给定的分多大资源和他的时间都是可控的，这样的话相当于一个砖的形状就固定了，对吧？

你包括所有的包括什么其实多大的矛盾，你和他的编程模型有啥区别？就是我们当时依赖于什么其实依赖他，我现在是一个现在在哭的编程模型上是一个happy的实现，不是虽然的model去实现，那个东西非常不稳定，而且知识也有限，所以但是理想情况他不应该是提供那一层提供更细层系列，多层就是允许一个sm上面的不同之一，就是你那些软件上面实现都可以变成他的to的那种时间，目前反正他假设就是说每个都是统购的，你的每个算法都是统购的，你要做一个必须要分支，这个东西做多了，其实对于指令的编译的大小非常大，看起来还是不太一样，但是咱们还是假设通过第三科的，我们是跟系里我们当时搞了一个刚好还是把他说上面不用他就。

说话人7 02:17:32  
不完全，他们人家十几个小时嘛，然后54我们的一个是的一，其实刚才有个可能是一个比较好的软件结合在讨论，没遇到边游，这样的话你是不是能够找到一个更合理的接口，当然讲你们的工作可以关注值得关注的。

然后另外一个不是讨论就算是概念，刚才大家基本比较清楚，对我理解。

说话人14 02:18:08  
接下来的圆扇子其实跟他那边的扇子好像是不一样的，我觉得不一样，就是说圆算子它是比如说我一个softmax是吧，它可以分解成几个步骤实现，对他是这段是把这几个步骤写成了元算子，比如说有一个map操作或者一个reduce操作或者什么的，听起来了，你这是一个算是在special上作了划分，比如说softmax我直接分成了几个小seb的Macs，觉得是纵向的分成几段，然后你说这样我觉得。

说话人17 02:18:40  
基本的原则还是逻辑上或者还是意义上，他这个是和硬件贴合，他是硬件的一个形象，就是一个硬件最小执行的一个对通过IP里面。

说话人7 02:18:51  
徐老师是通过胎儿里来实现划分，这个是不太好。

说话人17 02:18:59  
大流水流水就是算最核心的感情，然后很好做。

一个问题是这样，我现在思考有两个大的问题，目前我们经常讨论就是说两个关于这块两个大问题没有想清楚，一个是大模型的，一个就是稀疏，我一直觉得就是说但是从硬件角度应该是做成什么样子的，我现在的观点如果你做到这种任意的其实我规划早了还是不对的，你就不遵循你觉得发展一定做不出来，那么吸收最后的可能就会首先吸收很重要一定要做，其次说还是要做。

我刚说那种level就是说它最基本的朋友在level，在这个来我上面去做系数，这样的话就不会影响性了，首先就是因为你都是按照最基本单元去出版，其次就是说你要有个动态性，就是说有几个动态性，然后保证对国企的大量的压缩，然后所以这是我现在认为的观点，但是也不大正确了，因为我们现在他们也在试图做这个工作，但是好像规划的还有点太早期了，其实。

说话人7 02:20:05  
41年的情况叫2024，包括价格相关的一些东西，是是都不是钱，这样的话不因为也没有办法很好，你比如说您的发改委做这个结果应该在及意见，在国外的时候其实不应该说反正总的，因为我提出成立在这个方式去解决，但是有很大差别，不可能一个硬件全靠我，所以我做死了也不会，他其实很难专门拿一个硬件，就专辑这个是一定的程度的迅速度的问题，除非突然出现一个比如90%或者99%这边记录度，我再专门找他，这是很大的一个工程，除非是这样，我也想知道咱们从硬件层面去解析的这种方式，我就理解的什么东西，我可以解很那啥，我就加一个x其实简单来说就是数据配音next就是现在的吸收存储是反正我但是最清楚了，我们全包括想到自己来讲，他们就种不同的，我其实这个可以算是应该的。

但是问题是说我硬件一般不会说做特别通过专门对所做，我是个人感觉在短期之内搞更多的还是在编译的工作方面，或者说一些调整方面或者是这个层面去做，尽量把它变成成立的，或者说尽量怎么去做，我觉得可能应该就是你看一两年可能是这个思想，其实做一件也很关注，其实然后只是说我们现在拿不出4个过程是绝对的，但是现在从电力上面把这个产品的。

说话人13 02:21:57  
当然其实我在很久以前做过一些事情，但是最近两年没有跟过，所以我们这就是从编译层面做细说事情，他一般是是怎么合作。

说话人17 02:22:08  
我们今天做的工作就这样，最简单例子，比如说举这种你有几个点对你做细分，你不管做什么样的结果都很复杂，你们就看不到，我们现在就把它全部展开，暂时一条条指令，比如说乘以等于c对吧？

然后最后你变成一堆指标全部展开系数，你把这个指令干掉了，可是就剩下一堆指令指令的执行，其实它不符合数据了，都把数据都已经扣的 Data的， data call的扣子里面去了，这是一种办法说这是一种。

说话人7 02:22:39  
你得先拿到数据上的结构讲解，知道只是把他的Indies的数据放到致贫。

对，其实说反正计算其实都是假如以前我们出的这种数据结构因素实现了，从它因为主要是解决比如你落地的时候。

说话人17 02:23:05  
最后计数法线他们的吸收都是拿算除掉掉多少之后，你认为包括降了多少，但在实际中你的性能是不了，对，大家都说现在就问题是你怎么给你这个系数在GPU里面这些。

说话人13 02:23:19  
对我知道对，这样子的话就是说因为你需要先拿到这个项目结构吗？

说话人7 02:23:25  
就是说说的话相对确定的你发现现在就的使用支持可能，而且我觉得确实应该有戏的，除了大力度的系数为一的，我就值得做，那就算其实我们确定也是因为两个就是说还是有可能会对精度，因为其实还是觉得精度，你哪怕多算点多发点工号，多发点钱，还是要精度高。

九九百次就相当于是天天还是要重新的去可能对就是落地。

说话人17 02:24:03  
所以我就想知道大model怎么落地这个事情，这是我肯定还是要压缩的。

什么destination内容对现在好像听起来对这类似，然后再加上刚才就想，但是比如说他们做了那么几百 b的这种model怎么弄，我觉得这是很难的，对这也是个研究课题他这个方向错了还是这么走。

说话人14 02:24:25  
还是说他应该跟西书也有一些关系，就是说现在是不是有人发现了大模型，虽然有1000亿参数里面真正work的是一个非常小的子网络，有个叫什么彩票假说是吧？

他们说就是说他是那么即使CN里面或者是一个mlc里面，他们也学到了里面是一个子网络所以说，这是为什么压缩可行的，所以像肯定网络学得出来， GDP这种模型也需要压缩之后，要么它不是有两种，一种是结构化的，一种非结构化，你压出来以后就是没规律了，那就是那种东西其实在疾病上做或者加速器上做完全没好处了，就应该在c屏上就在CPU上部署就好了，结构化的压缩它基本上是符合你刚才说那种一块一块的，还是有一定的局部来说是的。

说话人7 02:25:19  
但是在宏观上是都和他们那个是才的，忘了是哪边没法，上海很明显的评价上市的时候，对开始调查情况下这个问题也是还没有说不同一个模型，现在注意他问题可以把我们写下来，随机然后直接上去处理的话，会不会这样子原来没有什么可以去 action到这个工作，就把它从带两个字用这样的规则把看看大家规定上，然后在面上去处理 blocks的，其实我是觉得如果只是在推理阶段去用把的大的把这种算法就有必要通过一个先进了，然后用的他们的干部在后面解决项目经费，有存储参数上看，对，但其实你可以通过压缩。

说话人17 02:26:22  
对有些观点就说不做，对就做那个模型的dhv类型加上可能对规则交流一下就可以了。

说话人13 02:26:31  
对，因为其实新说这个事情我们之前反正想过，包括智能网卡，因为我们当时是还在CV领域，我们当时觉得其实是一个伪命题，因为你知道。所以能系数是因为你的模型没信号。

说话人17 02:26:45  
还是有对一派人士的观点，对。

说话人13 02:26:50  
对，所以但是这是CV里面我不知道是别的领域是不是比如说像他们有些推荐哪那些旅游，他们天生就是个系统的东西，知道是不是？但我不知道是反正像崔健那个点他是不是必须得有什么现在很多。

说话人7 02:27:07  
然后所以一些问题是不是我们这样短期之内专用再去处理很灵活的技术，我觉得对结构化技术是比较简单的，没错，关键是这是一个需要去做的事情，是不是MV都做了肯定是可能的。

对，但是真的有人用起来了，只要能做了肯定就是觉得是这样是吧？他可能也所得出的，每个他能所在产品和家长这个东西，他的管理和各种见到的东西，他见到的东西太所以那个时候没什么对。而另外一个角度就说这个东西确实应该整改的，所以我觉得娜娜所以那就是见过他既然做了，他这个东西现在我把我们的串通，他对我的意思应该是不行。

就这点这方面的就出来以后你就是这4个中间2个这种事就。

说话人17 02:28:24  
应该的分享到这里，但是基本上实际上我知道的那部消息是他晚上效果很差，就在十几左右之前加速1.5倍，然后最新的所以他们做这个人，他其实我认为和比布格拉斯的快一点，而后来就比单子后来不拉丝忘了升级版本就变成200了。

说话人7 02:28:46  
所以其实这个还挺奇怪的，这东西都没有什么东西，但是好像是对我具体不知道，是因为我不太好找到拍板，是吧？

这样所以这个东西的话真正的应用领域的案例，因为这个东西所以这个东西我觉得还是比较特别像早期比较偏research的一个方法。

说话人14 02:29:10  
像深度学习不搞便宜也好，管线也好，是不是就放弃掉，你像CPU那么灵活，干脆只考虑一个局部的问题，就是现在最常见的这种case，最常见的算子，比如说吸收动态这些东西是就就用CPU来搞就行了。

如果是你完全不加约束，不加限制的话，跟以前那种CPU的对吧？写的用入的方式去写代码都一样的是吧？

说话人13 02:29:45  
对，其实我觉得动态的话，反正从我们的感觉来看还是一个必须去做的事情。

对，因为我觉得到的都。

说话人26 02:29:57  
太谢谢好的东西。

说话人7 02:30:02  
对，因为它只能说你可能就可以游戏可以去做，但是实际上你还是比较大的比较如何进入中间。

说话人26 02:30:08  
我的观察这里面有两派人两派人就说那种。

说话人17 02:30:14  
你做了model的落地的，当你要落地的时候，他的胆那里可能有像今天商汤不是讲整理很多动态系统，就是说动态性它是相对来说比较规则的一些动态性，或者是没那么复杂，就这个动态性我觉得就在基本上面去解决，或者是引起还是能解决掉的，恐怕是最好的。

还有一派人就是那种玩算法的，他就是想吃各种体系规范的这种模型，这种动态性我觉得解决不了，最后可能只能退化到做不好他就在Python乱装那些动态性，然后它相当于是你去一共更加flash的模型，但是性能肯定你写不好就很差。

我也能看到很多做算法，刚才写他想试 Idea的时候，他写出来那个东西真的很惨，但是这不注意到这里80%时间有可能，但是不重要，反正他就多跑两天就行了，然后最后然后在这出了结果之后，如果好。

说话人26 02:31:09  
再把它整个一个这是一个的这种 CPU。

说话人8 02:31:16  
加向量扩展，这种方式一定会比机构的这种 CPU加之评查就是一个其实比如我们用CPU加下扩展22+2t1或者2+g five加对吧对对这种一定会比我。

说话人17 02:31:36  
觉得我可能不一样，就像v的现在不是要做那个20家，但它不核心的区别在于不是水平向量和test的区别，它俩的处理的问题是完全不一样的，因为他们的对于法的OA的就是 opposition intensity是不一样的，向量只能处理的基本是以前，比如说最简单就MV视觉成像的但举证需要，除非觉得CPU加了这个了，靠他对内存的带宽需求很高，CPU的一般内存带宽很低，所以你就要做很多个窗口对吧？

也是做了几百个上千个，这样的话其实就没得。

说话人26 02:32:20  
最近要做一个什么样的话，就像我之前尝试。

说话人8 02:32:27  
了一把从MR的点击操作，然后往下优化，基本上大多数都是把它变成均衡，然后完了之后后面就是说放在中间点，然后我试了一把用向量，我不把不转换也不做负电话，也不做转换，就用一个算法把它先pass一下，然后两边再怎么着，然后但我试的效果其实还是蛮可以的。

说话人17 02:32:54  
对可以，你能打满CPU的我相信一定可以的，因为CPU的 memory和计算能力是非常低的，你要是在GP上面你肯定是比如同比的同等的工号或者其中的LV，你肯定打不过今天你光做客户录像这是一定可以，就是因为他的就相当于你每次从利润或者到网上面loa多少data的服务程度什么程度，所以他猜测为什么能做到那么几百100多tee，是因为我这里讲然后这不可能对不对？

然后我觉得这也是这一起的，但是你给我只做MV，就一份是什么GP，就没有这个考虑，你问他思考肯定做不过你，就是没有那么大的便宜了。

你用那个做完 CPA加上肯定告诉我们，就让他这个人事上面基本上然后跟以前合作可能大家都可以其中然后有7个亿，那么这一块这方式这块我感觉到什么关系到他们去，什么叫他们去，就是说老师我们就在做事，对转了一圈对我们因为业绩我觉得在小规模上面就是你单卡别的指标，但你在多少分布式的情况下大规模就持续了，我觉得核心就别抢着解决3000万对吧？

让他们去这些东西去表示，我觉得让他们就是说大家可以解决的，我们就没有没有。

说话人7 02:34:17  
但是达不到的信息，然后动态的就短期的东西。

所以为什么把他进来抽下，然后把一些东西给尽可能模拟出来，然后要让他们就只有在压力，比如说面市，比如说他这个事情的话转到然后。

说话人30 02:34:37  
或者什么大概没事。

说话人7 02:34:38  
我看了我们要求他们覆盖他们几个可能的分支都给他升职了，到底要看的时候他们都不知道他不要写，然后弄完以后大家就不需要他们看到的东西，他们反正他在本来就是说，所以说有没有东西这一块的东西，然后把他们解出来，然后他们出现一两个话题应该是先是然后讲出来之后。

说话人17 02:35:03  
这里面有个力度问题，如果你的 svo分支就是你的 svo分支里面的包的时候太小，因为它 igo GPU的东西它就涉及到了没可能，但是我说的非常可以，所以你一定要把这个里面的body做的足够大，就里面body是静态的，然后分支在外面是可以通过银行到这个就可以了，这个可能就是比较好的形势，如果你想完全做到你所有的都是心理上，这个平台，我现在然后我就。

说话人30 02:35:36  
肯定是在进行，

说话人9 02:35:38  
然后这东西那些后来这回你可能就说是怎么样哪些情况。

说话人30 02:35:45  
其实现在。

说话人14 02:35:47  
基本上像这种placement，就是说每个op在哪里，然后他每个阶段分配给他多少资源，就是资源的配额，这些事都是静态的，比如说其实每个AP它在运行过程中使用的都是Delhi开的，所以说在万福浪里面是这样，它的不是像内存池那样用过来以后它还回去了，它那块它一直它占着，但是具体说动态的，比如说这一块里面它用多少，这是run time的时候，比如说举例子就是这样的话。

说话人17 02:36:22  
你这样的话内存池刚才还是分好了，确定，相当于你要求这一个萨布瓜的整个执行的o而言，确定是。

说话人14 02:36:33  
是你们会确定顺像内存复用的，内存共享那个阶段是一定要有做 order。

说话人17 02:36:39  
对 word必须得确定这个东西，这块就是你在单卡上面，如果有并发的pass就没法做，什么不太好做。

就是说 tf的这种也可以。

说话人14 02:36:55  
比如说你只要说这多个stream它们加起来的，仍然现存的capacity仍然允许他们同时如果说如果你加起来超过了最大的，其实他连启动的时候。

说话人17 02:37:08  
其实当然是问题的，相当于最后直接是不能复印的，是那也是这么做的，所以我们因为经常编译也是要求把 memory提前分配好，编的时候把所有的AP他该从哪个地址访问到哪，就写全部都已经确定好了，提前算好了，运行的时候只要有特别大的出来。

说话人14 02:37:31  
它就全部是但是除非如果要发生流水线的话，就double8分或者甚至还有更多的8分，具体某一个半数据过来的时候，它里面反正最大配额是给它固定了，但是它是用这个配额里面可能分成三段，它用哪一段，这是动态的时候调了这样的，反正流水的时候，因为你可能两块它都在用是不是？

是吧？就是这样交替的话，其实我在芯片上有一个特别那种 aggressive的想法，不知道内容可行吗就是说大家都不知道。

现在其实做AI芯片里面，它里面做了几种器件对吧？有做矩阵层的，有做非线性变换的是吧？还有一些做向量计算的对吧？这个地方有没有可能是就是说它生产一种芯片芯片只做矩阵乘大小，就是我芯片里面很大的面积就是来做矩阵层的，全是做矩阵层的单位，还有一种是做分线性的什么的，但是当然你有通用需求的时候，你可以假如是一个主板上挂机挂着这种芯片挂着那种芯片，这个调度在这个调度的时候是不是会容易一些？

因为我们现在的因为你就像你用你每个东西分得一uw就非常独立，是基于对吧？你只要处理在这种大骗子之间的流动就好了。

说话人7 02:38:57  
好对那几个就可以，过一些三天，我一个鞋子里面只有一种光拿要么是举证成。

说话人14 02:39:07  
在现象变化，我们要明白为什么这种就对你至少在这个层面在宏观层面会。

说话人7 02:39:18  
它好像是一个分布式问题，就是这种model这个点，因为首先这样的话，我给我们看书的数据回去，最后大家都是偏爱还是偏偏偏外的话，不管的话一些什么速度什么都会很大。

所以一般来说我们不太可能这么说。

说话人14 02:39:40  
假如是流水线做的好的话，反正我不停的往这供数据，你这边反正这个是不停的去数据去计算。

说话人7 02:39:50  
不太会能把你经常去，本身他病情力度就很大，对别人先这个事了。

对这边都要这边的，比如说但是我们这10个他一开始是比较认真，然后是整机，然后是别的1个上市公司，然后这10个市场反正他们都没关系了，就是这样的。

对所有的三次个矩阵都放到现在，然后我的选送到这边，所有的别这个前提是他们之间不能有。我们要数据上对这个力度有点太细了。

说话人17 02:40:34  
等于就是说可能就相当于要把很多划分支付能力，他们快去做，那边做了还要传回来，传回来还是因为要出版再进来？

说话人7 02:40:46  
四s消费端内部的数据如果生产端都在进行上，而且我放的时候我们在思想跟现在银行的微软做一个叫第三个位置，这个时候而已，我觉得他不是这个概念，对他关键不是这个力度的好处，就说今天把这个东西然后别的然后这个思想我们经常做的时候，我们之前把他自己的在一个对这片可以理解，他们去一大堆了，对要都比较全面，对大家会看到积极性，因为存在异议的话，我们当然高声的基本上了解了从时间上怎么来的，是知道你们我们整个网络它那个拿东西，你就把他讨论对。

我听你们讲，徐老师你们把一个扇子换成了很多小扇子，这个小扇子你们有设计什么特殊的ir来表示的 u那个问题，我觉得童工你问的就是说他们 rap工具没有考虑以后的问题，我倒觉得之前的工作还是那种因为如果你有一个igo的igo的uw的话，这种东西往往是专门针对某一个运算，就是某一个比重比较大的影响，就像矩阵船萨克尔是针对矩阵桥来做的，那么针对这种你可以做专门的优化，我理解是像原来这个工具它是可能更适合于那些小扇子，它并没有是一套什么很完整的方向，把它给易购的在 sm这边上去做一个记录的。

的计算想做我觉得他现在还是说的奖，一个你像你那说我所以我就没有这个问题，我觉得这个问题就可以了，没有，所以我说你像你以后的机构的像泰萨克那个东西，你专门去处理那种计算量比较大的计算密集型的那种扇子就可以了，他不需要和其他的小扇子去做这种并行性上的那种合并，它往往是要去做那种 consumer和 produce之间的那种合并，对吧？

现在100或者说维护和好好。好对。因为他那个是一个，当然就说3月份的180就建完了给他圈的。对，了解这个过程这一段话，这里的话的孩子有一个现在想落实这种，比如说用户是还想去完成这种开的太高了，就是说的一个东西的组织关系的是结果，然后这三个东西没有其他的一些的对位的分析，我们只是通过打标记把他给细分项细分到不同的这种应用上，然后这样去解决这个问题。

认识对就的数据今天的对，但是对的刚才我们说的这种方式先把它把所有的计算都流到片上去，然后再去分布该去。

说话人18 02:45:22  
Q5的去Q5上去，然后其他所有不去QQ的就打个标记跑到 Vector和scholar处理单元的 memory分子。

说话人7 02:45:32  
然后去q我这边的一个自动生成的，对，而且还会转一下，通过因为这块转成把卷积转成集成成本，然后你不去的他打个标记，他下一步再处理，对，然后需要做的就做向量化，然后不需要的就直接生成刷刷代码就行，你在这拍卖的是先做你这个模式，对，他都是在那还来，我从这去做。

对，他这里头也就和他这里头用的你是好的这种模式，用他的那个事，后来还有那个方法也是可以对我们前面接他们。

前边可以的是的。

说话人17 02:46:13  
因为我们暴露完这个就需要一个还是需要一个策略是吧？对你先解决是暴露出来的，对没办法，就是这一块。对这种感觉我觉得非常好，其实都没解决，我觉得挺好的一个。

说话人7 02:46:30  
至少我们将这个开始让他们就可以逃脱，比如说你易购程度一点点到底是到哪个点就做不了了。

说话人1 02:46:39  
各位组长我们。

说话人31 02:46:42  
要差不多。

说话人1 02:46:43  
要开始收敛了，如果你还在发散的话，现在开始收敛，然后把 Excel表或者可以先做起来，就30。

说话人31 02:46:51  
到时候说你纸记吗？

你只记也行，你上来讲别人肯定看不懂。

说话人7 02:47:02  
不是已经讨论时间也差不多可以在一本就你们开始收敛。

说话人9 02:47:11  
我前面这个问题我觉得其实我知道了，反正我们就随便想到什么就讨论什么，但也没有对把问题记下来，对问题我是记下来了。

说话人7 02:47:22  
还有刚才杨老师说的那个问题，为什么要用专门的一个芯片来做集成，这个事情就我我。

我自己的理解是感觉它这个东西可能对编译器来说，它会少做一些优化的话。

说话人17 02:47:40  
直接大力度直接放过去。

说话人7 02:47:42  
对，就这样的话其实编译器的压力会小一些。因为我觉得就是说你大力度放在一个基础上，我们可以放一个拍摄特点啊，我不可能我一个公司我为了做一个我做欺负来做，咱实际上就是好多人什么就是新店我觉得会简单。

不你说我这个设计的时候，这上面只有集成厂，只有飞行电话，这个不是设计的一个挑战，这个不是说你可以说我这里那块面积都共举了成做一个很大的事情，然后后面这个 KTV就这样了，然后你后边做非常清楚的人在另外做一大块就在一个芯片上，这不是问题，就是你说一个芯片光做这个东西，他没有讲话，我们基本没有讲。

想到我的工作，其实就是说我觉得你可能是一个问题，说从编辑的角度来讲对吧？你是一个大的军人的东西你好做，还是说一个什么，我以前理解这种难度，我觉得tpo的你用起来不是难吗？也难用，就用不了那么大的数据，你要有效利用它不是难吗？我们还没太明白明白明白，你说是你说少于。

说话人18 02:49:05  
比如说我们在一个片上一个芯片上，它可能有不同的那个功能部件，我先要把所有的数据都搬到芯片上去做，对吧？

然后你又要把把我们先说算算子，我们刚才说算子到底有没有意义，我们先不讨论这个问题。

你这几个算式之间你要做future，我先要做future，然后上去之后又要分出去，其实这两个过程是两个相互相就是相反的一个过程，那么我如果没有芯片上不同的功能部件的话，本身是一个芯片直接插上来，然后然后直接少了一层 Memory。

说话人7 02:49:50  
这个层级可能就会少了一层，这样的话我这一步的修正可能都不用做。

直接我像我们当时针对。

说话人18 02:50:02  
 ahhhhh生成芯片的时候，第一层就有一个LV和所谓的Eli和 ud这样一个方法，又必须unified，对unified还是支持 vector还是干了的。然后你先要进到芯片的芯片，整个芯片的访问的那种buffer上去，然后再分到ewb和Eli上去，现在可能Eli和ub就没有了。

所有的数据可能。

说话人17 02:50:27  
都在这样，没有数据的没有没有没有，还是你的数据库可能上不去。

说话人18 02:50:31  
对就是说我的意思是说刚才袁老师说那个意思可能就是一个芯片，就只处理这样一个功能，然后数据这个事情，然后就不知道硬件就怎么解决了，然后我的意思是说你没有这层级的话，然后编辑优化的时候可能是会好做一些。

但速度这个事情我也想问一年他比那个季节。

说话人17 02:50:59  
你这样就可以了，其实就是tbo第一代就是这样。

说话人14 02:51:03  
他就是一个大巴是一个大的还不够，就是说所以从分布上来说也可能有一些好处，因为你现在一个transformer我是做了一层计算，然后要做下一层计算，它实际上是如果这一层很大，它是已经分到很多卡上去了，你说这个矩阵能分到很多卡上要做下一个计算的时候，这些数据又要收回来，然后才能做下一个计算是吧？

如果我这一个矩阵单元只做一个就我我也不分了，直接让他在这个做了，做完之后一体的一把就传到下一个大的计算单元上去了。

说话人7 02:51:39  
他没有中间的很多卡之间说你们未满这一个大的我就没问题了，我就给你打了，可能大的模型上会有好多，因为它本身现在的问题是就容易被马元做是大大的那那，多个任务共享，对。

就工作可能还行，反正中国还有没有其他人的，就是我们这卡担心我们这算挺大的，我们真担心是用不起来。

说话人17 02:52:06  
又不现在除了传统的系统。

对。

说话人7 02:52:10  
所以我的问题就是说咱们去做这些igo的话，是不是真的有必要是吧？比如说我觉得其实在科学的方法，可能还真是咱们最初的刚才他们先建一个这种只够加一点 Igo的东西，然后我们先去看在学校里面你可能我发现你这个以后这个问题很忙，忙得过来你还得它可能增加一定程度，我觉得这是一个科学的去从工作上去，只能得到这个架构的一个方法，对吧？

你可以先开始做，我这就是个所以说4×4对吧？我可能觉得或者怎么加，这个比例不就是一个科学的得利，所以我觉得一共有一件事被逼无奈，合作一个可以这么说一会儿，不然的话，因为你毕竟一旦很多时候，一个当然做芯片就是他没法说，因为你来工作我做的你可能44×4，我再来一个就变成不对了对吧？我肯定是找一个平衡或者力度细的或者是一个平点对吧？

但我比如跑上一一百种logo，我在中间找一个比较合适的点也行，反正至少我觉得这至少是一个我们说的定量分析的话，我们不是纯凭经验或者是一个分析方法，我觉得这个这个其实我们也是想做，可能和你的是这之前的思路不一样，因为我们想联系的话，我们是没有现成编辑，我们想用MR作为一个现成编辑，但是这个思路挺好的公司，我觉得可能所以你们得往后加一步更容易一点，但是我觉得我无论是原来是因为看到了有一些现在的架构满足不了，就是说。

说话人14 02:53:58  
但是从分布式角度来说，我觉得反正是比较大型的模型，用那种我觉得会就是说比如说什么能效比也好，或者是我觉得会更好一些，他现在是我这个数据进到一个算字里面去，这个算字就会分到各处去了，各处去一定要交换，才能进入下一个算式，看到每一步都要经过一个计算交换计算交换，但是如果你你算期限足够大，我举证成我也不需要怎么划分了，就进来就举证成，成完之后整体的输出就一条通道传到下一个什么，反正就复杂的要解决好就是路由的问题。

说话人7 02:54:41  
如果说你最后大模型或者什么用的是不多，或者很那个的话，也许会出这样的一个结果，对吧？

专门针对性的做这么一个硬件，这有可能，因为其实从角度来讲，我可以做一个很大的，只是我们不知道他用在哪里啊，比如说假设你一个a版的芯片的面积，我全做全程都可以做很大，然后可以说他要不是他好说老师他只是说。

说话人17 02:55:08  
可能比如说两个比较，我觉得这个也符合我们要讨论的这种方法就是说，因为大模型也是一个新的出现的趋势，到底做大不知道，所以我觉得他这边就分析，有一天我们发现大的空间可能占的比较多，对，所以呢你就会推出来，我们

说话人7 02:55:26  
不在这个家，我们其实就想知道这几个东西对计算和主动性通信这次占的面积好了。对吧？我们下一步最大的功耗的限制在限制上它这个比例是不合适，其实我们就想知道你这个问题，我们知道我们就可以写芯片，我们就比较大的一样的设计平衡，其他都是在这个上面再去就一些专业专题规划，再加上然后相对来说8.989 9 13 13，现在哪个时候这个肯定其实大家都是比例不一样的，但是的话每个人做出来，然后我们就看咱们有了差不多，然后如果第一个能不能找到一个合理的比例，对吧？

或者至少科学的去找人，所以我不知道就像刚才那个地方说的，你们是有配合，其实还是说新东西都还没出来，对不对？

在是不太理想，但是前天晚上在烧烤的在那里，然后到了5:00的时候，我后来决定放弃了，我觉得很多地方不太理想，所以我然后就说当时那天晚上也是想到了很多当时黑板上没想到的东西，然后重新做一下算了，现在计划到了，所以你认为他没有什么感觉，但是简单的是不是你看他损失定50块钱，创新这一块我觉得还是挺重要的，帮助去探索这种东西，对。

对，我一开始看我其实就看到他，我没有看到他的，我先不关注他端到端的变异，确实是就像刚才谁说的，不知道砸到哪，也许是你想不到的一个地方，他好多时候他去上路还可以，因为我现在估计不知道什么编译速度什么这种问题，但是我至少上手一会儿。

还挺快，对至少能够让大家能够能赶紧干活来，对，就说他推出这样什么这些东西，毕竟还是给你一些工具在哪，要不然你自己写c加，就不需要这个地方，其实我不知道除了像 Mr或者刚才其中说的这种，我不知道还有没有别的工具啊，就是说从咱们可以帮助咱们去加速我们的一些可以探索的方式，我刚才说的他们一个感觉从一些架构那个东西不看不是很好的，比如说一个东西咱回头可以聊一聊，他有叫芯片设计架构探讨，因为其实我也不只是去愿意，所以其实我。

说话人13 02:58:06  
也不是做便宜，我不知道可能是张老师是不是做编辑的是吧？

对我不知道。

所以从编译的领域，它有没有一些那种用的比较多的这种工具帮助去开发或者设计编译器的这种？

说话人18 02:58:25  
我们解这。

说话人7 02:58:28  
放心用的比较少。他们会比较少，对，会有一些辅助的交互式的那种工具，前端的工具其实比较多，前端的工具对前端的会多一些，对

说话人18 02:58:50  
我们自己做优化的时候就是有一个优化的工具，它会做一些用的 ISL的教师工具叫icc好，你自己可以在里面验证手动改一些你自己优化的这种表示形式，然后在那里面验证一下，然后验证之后我们再写到编译器里面去，但这种感觉很少。

说话人17 02:59:16  
对，首先是我还有个问题就跟发现一个问题，就说今天谁讲过很久还是大model，我很纠结，就是因为这个大model只有大公司才能确认，然后我们该怎么办，就是说有一种思路说给我们做一个民主化的确定大波段，因为大波段一定要吹出来，它的泡沫非常强，它掌握了所有的信息对吧？

基本所以就很强，比如说我们做一个完全分布式的一种缺点，这种分布式就是到那种去中心化，每个人给他贡献一些GPU，然后你就加入这个节点，然后去群里有可能会慢一些，但是规模非常大，大家所有人都可以加入，然后就像那些练一下，你加入你去吹对吧？最后大家一起去吹出来，一人分一部分 share对吧？

然后去做，就这种power我不知道能不能有可能会对抗，就大公司它因为它有一个非常大的class能做出来，否则我们平民怎么能做出来，我不知道这样一个方法能不能做出来，比如说有什么差旅，然后我是觉得很蛮有意思，就是发动我们去推整个社区去弄这个有可能弄的比GPS更大一些。

说话人7 03:00:28  
你们只考虑过跨机房或者什么贷款的，以前搞我有台电脑，其实现在也在那传统以前有一些网格计算，对联邦学习是可以用的就是。

说话人17 03:00:44  
说但是你需要一种机制把它组织起来，最后 model出来之后，还要而且大家需要快点对所有人都能用。

对这个。

说话人13 03:00:51  
应该是有挺多研究的，就是他们做联盟学习的怎么样去激励大家。

说话人17 03:00:55  
其实我觉得是可以的，对对就是我们创造一个活动。

说话人31 03:01:00  
然后大家想挖矿，

说话人17 03:01:04  
对这玩意他基本上是可以是行的，没有人做我觉得是南向联邦学习是证明了可以很难给我去做一些事情，但是我不知道去大model参与的是不是？

这个真是需要很多人大模型需要交换的数据太多了，是的对。

说话人31 03:01:20  
需要一个铝word，就是你加入挖矿了你要得到一个回报，然后你才能进这个就是。

说话人17 03:01:27  
你最后得到 model的一部分share，将来 model可以用商业价值非常强的，他卖一次一片Kao分给所有人，现在。

说话人9 03:01:36  
我这知道他像现在串出来的大摩托，他就给你的API掉，他最后你做这一次EMS是他是多大的，我的好说给。

说话人7 03:01:49  
你的节目就行了，这个事就是没有保证的对吧？

你也没法验证的，就是技术实际上大部分软件也。

说话人9 03:01:59  
是搞了一个大的，然后搞了API的。

对真的是都快s了，扣子扣子扣子拍的对，我比如假如说现在我调出来一个，然后我写了一句话他给我补完了这句话，他实际是 us的过程，它是也用了很多资源，还是说还好，就一个靠的话是还行，还好的话，因为我们现在也是顺着大的工作，他做过effort，难道不是整个毛都要动吗？

还是说你对你们也是放在CPU上，网其实放上，就是说你一次一本的话。

说话人17 03:02:38  
他的核心是说真的太大了，还有每个一段时间的时间了，对，但是他一段时间太多了，因为过完一个一套就是要做，其实上他们确认的model大部分两个一块关键是几个月就结束了，我的理解这样，然后那么其实你一会儿再靠自己跟你那种人脸识别的靠的还是打的那么多。

说话人9 03:02:59  
就在这个东西的技术做也要费很多资源是计算很多举证成就相对对象一遍也得。

说话人17 03:03:11  
你不要说了，关键这玩意儿他的能力太大了，如果说他数据都说起来吹，他都记得。

说话人9 03:03:17  
我理解他就是刚才咱们说的这些产品来说，只能中间只激活中间一部分啊或者这些手段，最后导致应用是可以带上小一点，真的还是可行，要不然也他还是关心我觉得我不知道他怎么做，我觉得所以现在只能做成很多service，他毛总没办法做到端上面，所以他就是4人所我都不太能理解，就是说你那么多人，我这个光做优势总觉得其实也是，除非他觉得这个讲话的。

说话人7 03:03:50  
不可能，我说你刚才我跟你我一个现在一般价格我给你准备好行了，因为我记得那东西我就说算了。

说话人1 03:04:00  
好的，两位上面好我们时间差不多了，我们现在下面这个组就是小组代表发言，然后我们先这一组第一组啊，

说话人7 03:04:13  
就是因为我记得我在纸上记得可能其实我看清楚了，所以我再聊一下，我们其实讨论了几个问题，有些没答案，有些好像有点有意思的答案。

第一个我们在讨论纪龙这边刚才讲的，他们专门往下的一个进呃，下一步的事，他们可能是想他们是希望说把硬件做一个抽象，刚才他其实那个图大概讲了一下，我觉得我之前一直对专门有一个疑问，就是说专门它对硬件的假设这个力度是一个同构的东西，它每一个优势同构的东西，面对我们一些至少。

说话人31 03:04:54  
你的 sm号。

说话人18 03:04:57  
加上参数call。

说话人31 03:04:59  
这个。

说话人7 03:04:59  
东西它。

说话人22 03:05:00  
就不好解决了，对吧？

但是他们现在的工作可能往后。

说话人14 03:05:04  
会把这个村村号加上。

说话人7 03:05:06  
然后当然相对来说，它还是对硬件有这么一个抽象，或者有一个建模，还是一个简化的模型。

通过这个模型。

说话人1 03:05:18  
我的理解因为我觉得不对什么，然后通过简化的模型导致编译器会。

说话人14 03:05:25  
相对来说工作会有一些相对来说容易做一些强调它的一个基础。

然后我们讨论这个东西，我们讨论东西发现他有一个有意思的地方就是说，他有可能会得到一个我们一直在说一个合理的硬件架构，或者是一个。

说话人3 03:05:44  
从编戏比较友好的硬件的架构是什么样，我觉得虽然现在他可能刚开始的建模还是呃是一个比较简单的建模，但是他有可能回答一些问题，比如说。

说话人7 03:05:58  
他现在的思政炮加上成色套，那成色套我可以先讲一个比如是4×4×4的这么一个层次号，它 word下来编译以后可能会有一个结果对吧？

可能会发现什么枪支票很忙，我是不是枪支票再这么大一点，对吧？

因为它是个硬件的模型，或者它是一个从编辑角度来看硬件模型，我觉得这似乎可以往后走去看一下这种，其实我们做硬件是设计这些手段。

说话人14 03:06:29  
我上午也讲了，我们这些可以用什么手段，其实最终就是面积用在哪。

说话人3 03:06:36  
对吧？面积用在哪种东西上。

说话人22 03:06:38  
划分对吧？是用在 General的。

说话人7 03:06:42  
 process上，是用在成熟call上，还是用在 RAM上，还是用在 communication上，对吧这？

说话人14 03:06:48  
其实就是个比例问题。

说话人7 03:06:50  
我们其实就想知道哪个和比例对于我们常规的logo是合理的，所以我觉得他这个工作当然一个是编译的工作，另外一个是可以往后走一走，看看软硬件。

说话人22 03:07:01  
一个更合理的硬件更合理的一个样子。

说话人1 03:07:06  
我觉得这个挺有意思的，这是我们讨论第一个问题所以，他们说你一句话能总结这个问题，我不知道金融他们后面的工作叫什么。

说话人7 03:07:18  
比如叫软木二对吧？

二基于软木二的。

说话人3 03:07:24  
软硬件这个叫什么软硬件接口，或者说我看的时候。

说话人1 03:07:32  
软硬件接口抽象。

说话人3 03:07:34  
之类的，或者说。

说话人22 03:07:37  
再往后合理的软件划分，我觉得他可能我们一直在谈软件给硬件给软件提供什么样的接口之类的，这是一种。

说话人1 03:07:50  
探索，我们觉得这是个problem，这是吧？我觉得这本身是吧。

说话人24 03:07:56  
问题可以叫成合理的软硬件接口是什么对吧？

说话人7 03:07:59  
对合理的软硬件接口，对通过专门为to去解决，对从边界的角度去探索一下这个问题。

说话人1 03:08:08  
所以回头我们再抽象一下，我先讲就不占大家太多时间。第二个我们。

说话人4 03:08:15  
我就刚才说了一下这个算子概念。

说话人1 03:08:17  
到底有没有意义，大家也简单讨论一下，反正这可能从这个算法逻辑概念上，它是从逻辑概念上来的，但是现在其实用的中间。

说话人22 03:08:30  
的包括算子的边界，但是包括融合，它其实也不是很强的一个。

说话人7 03:08:34  
一个概念了，反正我们大概讨论的是这个结果。

说话人1 03:08:39  
然后另外一个我们也还花了一些时间讨论习俗化的问题，就是啊就是其实说话的方向到底硬件。

说话人5 03:08:50  
该做到什么程度，怎么支持这个习俗话。

说话人1 03:08:56  
这块反正我们对可能也没有统一的结论，只是说反正硬件要支持不同程度的信息化很难。

说话人14 03:09:10  
其实可能现在短期也就是这种结构化的提出先做一下呃，可但是编译器在中间可以发挥的。

说话人7 03:09:17  
作用应该是值得讨论的，可可做的事情可能是比较多的，包括一些新的数据结构什么之类的。

说话人1 03:09:23  
这是大概的结论。

然后另外一个讨论的就是说。

说话人14 03:09:28  
这个也没结论，就是说大的model今天讲了，上午也讲了，林总也讲了，大的model到底是。

说话人1 03:09:38  
我们老百姓怎么用。

说话人14 03:09:42  
有点比较不让storm的一些讨论，包括是不是我们我们按这种网格计算的这种分布式或者联邦学习的模式，我们每人贡献一点这个东西我们能训的，大家一块训一个用自己的电脑组合起来训一个大的帽子有没有可行性。

说话人5 03:10:00  
然后包括他。

说话人7 03:10:02  
去完以后他怎么落地，就是他这种很大的model。

说话人1 03:10:06  
他给你用的时候，现在可能大家有时候给你提供API什么的。

说话人7 03:10:11  
它背后是会做一些压缩，还是会做一些 dcl什么等等这些没答案，我们也不知道各位如果有答案也可以给我们讲一下。

说话人1 03:10:22  
然后下再一个问题就是说学习。

说话人14 03:10:27  
讨论了一下。

说话人3 03:10:29  
动态性和性能的问题。

说话人1 03:10:33  
这个咱们刚才是什么动态性来着，我都有点记不清了。还是税务的动态吗？

反正我们好像也没太明确的结论，就是说可能也还是吹到什么问题，可能这边有结论。

好好。

说话人14 03:10:54  
然后刚才又提了一个新的问题，就是我们今天可能说compiler比较多，煤炭说 long time的问题，就是说这个状态就是Conde和状态，我们后面是不是有一些更合理的分工，或者在某些情况下什么东西放在什么地方做什么，我觉得今天没太讨论，我们也没什么结论，提出来这个问题后面也可以大家有兴趣可以再聊一下。

然后其他就没有什么太多。

说话人1 03:11:29  
更深的，我们其实就是第第一点是讨花时间讨论挺长的，觉得也挺有意思，后面可能会。

说话人7 03:11:36  
也期待这边一个工作，我们也看出有些合作在里面。

说话人1 03:11:41  
要记住要补充一下 rem指出这个结论能不能用更加通俗上的话来。

说话人9 03:11:47  
来说一下，因为rap可能我们就想出现以后叫做cancer的baby。

说话人17 03:11:54  
然后试图来通过比开始准备作为一个软硬件的接口，这样的话就不上了。

说话人1 03:12:00  
对就可以，所以跟现在的这种有什么区别，跟现在的这种软硬接口有什么区别？它变更简单了，更加单一了一层。

说话人17 03:12:08  
还是对它的简单就是对硬件来说比较实现这些方面的，能够然后对软件来说，我们的编译其实很大的重要意义是可以做到百度编译，就说是以前的编译像TM和你搜索这种去做，你很难理解他的行为以及编译的时间很长，然后白鹤编译的话就是说你可以一键编译，并且系统的可以做分析。

说话人1 03:12:29  
相当于有了cost model，大家都提供了对好的大概该到的意思了。

那么第二个算子概念有意义吗？我记得敏捷是不是也有问过这个问题，其他两个讨论有没有讨论过算子的概念？这个。

说话人7 03:12:46  
这个是我们还是没有也在有再有大概几个点，一个是说算更多的是一个研究员的一个概念是用来做交流，还有一些习惯，但是在真的算的优化层面不是非常必要的，然后。

说话人24 03:13:07  
现在大家都在做阻断融合不也就是在打破边界，然后。

说话人7 03:13:14  
后来我们还探讨了一下，就是说我们假设不走这个层次，我们能不能有一个比算词更往下的，但是能够相对比较统一的艾尔表示，然后大目前看起来很难，所以你刚才说更统一的比在上次往下更统一的，他表示跟那边而且有点相似，这边的可能就这边一种回答的思路，对，至于能不能统一我们不知道。

对就也许可以，但是现在明显还没有一个足够标准的东西，它不像大家都看过都是一样的。

你们有什么讨论吗？咱们有可能讨论这个问题吗？就是讨论出相对来说大家比较有共识的，不是什么是什么关系，你说的这方面不是opera，但我再往下，对于不管是编译还是硬件来说。

说话人8 03:14:08  
都更友好一点的这种，你刚才不是这个问题，是也想往下讨论一下吗？我不知道你们讨论，我们就停在这儿，行，我们就走一个问题。

说话人1 03:14:21  
因为其实两个组其实都停留在这个问题是有共识的对吧？我们要在甚至下面可能会有一个更统一的因素，本身，就是刚才的意义我们应该是一样的。

说话人15 03:14:32  
思路意见一样。

说话人3 03:14:36  
好爱你远看还有什么其他的。

说话人7 03:14:39  
动态的什么动态的动态性这块，可能我们其实现在是我觉得咱们是说动态摄影的支持的意义是吧？

所以也有没有必要支持动作设计。你再改一下让你听一次是吧？如何支持我，这肯定有必要是有，你如何支持我们。

说话人13 03:14:59  
当时对我其实有没有变化，发大力气去做。

说话人7 03:15:03  
如何支持或者huo，那么就是说动态社会的价值到底有多大，然后如何支持是吧？这个问题对是不得不支持，不支持违法就必须要支持。这个问题能算下去这里。

说话人3 03:15:19  
可能两是两个问题，静态图是不是要支持动态修复，这是倒是个问题，但是你从这个这个这个模型的角度，你一些检测模型本来就是个动态系统的是吧？

当然你也可以去强制转化成一堆的动静态图去搞它，这个感觉太大了，所以说这里这个问题是不是说静态图里面要不要支持，刚才先我要理解一下这个问题，好像我们的问题还是说支持这个东西又保证性能。

说话人7 03:15:45  
他应该是途径是什么？还是说可能损失一些性能去更简单的去识别。

我觉得这个和性能之间的渠道防止这个问题不是指的动态compared就 Compared的时候动态看到来确实是。

说话人24 03:16:01  
还是是框架层面的动态还是。

说话人15 03:16:04  
编译一个三组的时候。

说话人7 03:16:11  
我记得我没记我也记不太清了，没有关系，这个关系先放着。

你你，你后面再改一下。

说话人1 03:16:23  
雪峰里面的结论就是说是这个问题重新定义，应该是说是否要动静态。

说话人3 03:16:30  
是否在静态图里面要去，我只是刚才跟跟他们欠了一下到底是什么问题。

说话人7 03:16:38  
我们就好像聊的多有点忘了，就到别的问题了。

说话人1 03:16:45  
好，那么我们可能还要让他们看看他的时候更合理的分工。

说话人10 03:16:51  
另外两个他们有没有一些。

说话人7 03:16:54  
有没有讨论过这个问题，有些高可能有些观点一些小问题。

说话人10 03:16:59  
好。

好，其实之前也简单讨论了，就是说要让他们看就是说因为我们确实今天主要考虑要看看，不过我不太清楚另外一个小组怎么定义这种，因为我我可能会更多聚焦在说我们的compile编译出来这种版本的文件以后，我们让他们要负责把 better比较人进来，然后这个时候说可能会跟芯片会很平的，芯片这种driver非常相关，就说你的driver到底是开放，比如说初刻力度或者吸客力度的接口，包括你的眉毛这种接口是不是这么开放出来，我觉得我可能更在乎的是讨论，然后他们是所以我们来清楚的是这个小组怎么定义是这个往上看。

说话人7 03:17:57  
我们也没讨论太细，我们还是说哪些事儿，比如说在让他们要干要更努力一些或者是什么，我们没有什么结论。

今天好像没有让他们跑的比较少，好，一抱一，然后定期来调查，你们要回应不要委托，所以没事，我们只是感觉今天让他们没太讨论，可能后面知道这个问题，可能后面。

说话人15 03:18:30  
可能值得讨论一下这个意思，或者要不你要补充一下，在防范这边。

说话人10 03:18:36  
其实我们是让他们的话，我觉得说可能我觉得如果从芯片角度来说的话，我们让他们怎么和芯片和Java更好的一个配合。

比如说芯片现在很多加速芯片是非常大出克米的，还是可能比如相对gpo或者怎样，他们会是更希律所，然后现在很多生态的环境状态确实也是更加适配西客利率，因为他会在软件层可以做更多的事情，这个时候怎么去做这个去落户，我觉得让他也是需要去考虑的。

说话人1 03:19:13  
好的好的，那么辛苦志愿者记一下，然后我们一下一位就是请杨军是吧？差不多已经记了，对好的，谢谢，然后我就把我们这些小组讨论一些问题，找一些跟刚才谈过分享的。

说话人18 03:19:36  
有些互补性的会稍微说一下，如果有些我有遗漏的，我觉得可能大家给他们补充一下。

说话人15 03:19:42  
但是第。

说话人5 03:19:43  
一个问题就是说我们今天讨论aabe的话，其实提到了这个软件系统设计，那么软件的背景同学跟硬件背景同学他们的外观是不一样的，在这个背景下怎么能够有一个更有效的协同设计的机制，包括怎么能够形成共同对话，可能是我们觉得在这种BA系统探索里面一直没有微信。

那么这里边已经说了一个事情，就是说如果我们能够有一些开放的可以共享的基础的话， Maybe能够简化复杂性。

举一个例子像real的话，其实它在一定程度上讲，简化了我们为一个新建设计Isa以及软件站的一些工作量。

说话人22 03:20:22  
那么他可能对于所有事情，我们发有的公司我们可以有蛮多可以事儿的东西。

说话人5 03:20:28  
甚至一些我们觉得可能相对不是那么的敏感的东西的话，可以贡献回去，相当于我们把整个社区的力量和公司量形成一个共赢。

那么这一层面，那么软件层面的话目前能看到可能还是各家有各家的特点，TV、mx or whatever，那么软件层面有没有可能也去类似的传出这样一个局来，然后这些不同的厂商能够往一个大家有一个共识的，一个软件基础里面一个共识软件靠的，恭喜我们理解，那么以后如果再有出来想做新的硬件的话，想做新的硬件软件，这样的话能够去拉入这些工作。

因为这是我们讨论的第一个可能觉得比较有意思的一个问题啊。

说话人1 03:21:08  
所以稍微再总结说，你是觉得像在软件这个站里面是不是也有像类似范围这样的一个。

说话人5 03:21:17  
联盟生态，对MR可能有希望在MR的出来在比较的晚，目前整个成熟度和完美度的话会差一些。

对，所以可能我们觉得还是另外m二社区本身也不是那么的凝聚，它缺乏一个像类似于天气这样的一个灵魂人物来统一教育这个方向，导致发展的比较的多样化。

对，所以软件上目前法人没有那么的成熟，在历史上对对。

说话人3 03:21:38  
这是第一个问题，对。

说话人5 03:21:42  
第二个问题其实是一个可能跟乌克兰有关的问题就是说。

我们到底在训练场景对于这个性能有多么强的需要，那么可能我们在场有很多做a系的同学，我们关注说20% 30%的其实对训练，那么从业务来说的话，可能很可能说百分之二三十的性能提升，不如说你给我更多灵活性更有帮助一些。那么在这一点上到底气能到什么程度，是我们做新的系统新的硬件效观测点。

当然这里面有一个有点意思的事情，就是当我们说系统的话，我们不可避免需要跟未来对标，那么NV跑的速度基本是说每2~3天一次大迭代，可能到了明年10月份的话，如果没有意外的话，他们的户口会出现性能可能是a版的三倍，也就是说哪怕我们不考虑性能，也需要考虑说不能比他差太多，才能去考虑应用性，所以这就是我们在新建公司里面的话不得面临的一个挑战，我觉得这是一个我觉得比较有趣的一个问题，对。然后可以好总结吗？

说话人1 03:22:45  
不太好，你不要没有找短一点的。

说话人7 03:22:50  
这是一个再小一点，一个是大的，那么先这样谢谢。

说话人5 03:22:56  
然后第三个问题是这种大模型其实主要是商汤的同学来贡献的一个观点，因为他们可能离我的客户更近一些，大模型的训练其实会发现我们每当换了一个model，我们需要在整个切分的策略做很多的一些工作来保证我们形成讨论更好。

可能不同的model切入效率不一样，因为标准可能荒漠的话你可能会好一些，如果再复杂一点可能不一样了，那么这个里边我们对VIP这种自动化的分布式处理的要求会更高一些，这个可能听到了我很开心，因为工作更强的需求，对，所以我其实觉得可能是一个题外话，我就觉得我们讨论应该更多的拉一些有这种比较强的沃克露的背景来参与进来，像上午林兰花老师分享，我就特别有启发，因为来自于客户的分享的话，能更好的在我们整个系统研发的方向，包括软件和硬件，这可能第三个的问题。

说话人10 03:23:58  
其他的。

说话人5 03:24:00  
其他的我觉得可能是一个相对小的点，就是因为有些问题在唐伯德里面也提到了，所以我这边就不再提了，我不知道小组里面有没有什么想补充的问题没有？

大家可以来帮我补充一下，也想去听他的topic。

说话人10 03:24:13  
一般哪个人。

说话人15 03:24:18  
报考好体验。

说话人10 03:24:21  
我一个就是说我们说是大家现在做很多AI系统，说比较分散，是不是有可能大家会更具体一些力量去做这个东西？

说话人15 03:24:36  
对其他我觉得。

说话人5 03:24:38  
这个主题其实中午吃饭是金辉也机会你还在机会，因为明天赶火车去了，对机会提到我说我们这么多人其实都是属于做这个领域的一些同行，其实都代表了不同的，视角如果能够有一个合适的策略，把大家以某种方式能够三个起来，可能是一种虚拟的 PPT什么，我待会能够去把这些对这个领域的理解形成一个common的可能会用这些东西，可能对于整个行业的进展，对于自己的这种比较有帮助，可能这是一个我觉得可能也是想蒋老师讲讲，到时候去传一个大的三年，可能希望慢慢去引导去促进的一个东西，一定是想具体的一个topic和问题，希望能持续讨论。

说话人31 03:25:19  
我们其实最后有准备一个环节，就是我们要组成讨论组一个小的肯定是吧，然后名字没起好。

说话人1 03:25:29  
反正叫一个执行委员会，然后我们来探讨到底。

说话人31 03:25:33  
我们要出什么样的一个report。

说话人1 03:25:35  
然后去组织这个会议对吧？

我们肯定会讨论行。没有的话我们就我刚才在 U盘里面有他们的问题，新人的手速是非常的快，我们敲很多，谢谢新人一看编码编的很有必要，见文本上面。

说话人7 03:26:02  
我们在公示，我还有每个人都已经打上标签了，然后然后。

说话人3 03:26:10  
因为那个内容比较多，我挑挑一些讲，然后可能有些大家觉得不全的补充，大家第一个就是说为什么有算这个层次已经讨论过了是吧？

刚才已经讨论过了，所以说我就不再赘述了。

第二个就是我们讨论了一下怎么构建一个更好的编辑的心态，实际上就涉及到编辑能不能标准化是吧？能不能把上面的复杂更好的跟就映射到硬件上这里面。我先说自己的观点，我先讲一下就是说边际生态标准化或者说开放化的当前面临的一个挑战嗯。我并不是说有说这个领域不能形成标准化，我只是说它目前存在的挑战，我觉得分两个维度。

说话人4 03:27:09  
就是说首先说我们。

说话人3 03:27:13  
谈 Mr，刚才蒋丽老师说我反对MR，其实我不反对m因为我们公司也有很多团队做MR只是说我现在对它做一个标准的 AR的编译器，可能我觉得他挑战蛮大的，第一个是说因为m二是一个分层的编译器是吧？

但是其实在m二出现之前，就是我们那种商用的公司IBM，包括它里面内部的编辑，它就是一个分层的编辑，但它的分成的编辑跟MR的分成编辑是不一样的，就是说MR的就是像比如IBM或者说我们华为内部也有一些类似的编辑，它的分层是定义好的，我是两层还是三层对吧？

比如说IBM的编辑，它第一层就是一个类似于烂的一个表达，就是它的huo这种东西，然后到下面就变成可进这样的一个但是 m二它是m二它实际上是一个机制，它告诉你了我可以做分成，但是他没告诉你你有几层每一层是啥，所以说它不是一个标准，它是一种公共的机制，所以说我要澄清一下。

第二个我们也要要理解，就是当初lvm为啥会形成一个比较好的编辑的标准是吧？那是因为我们在芯片硬件层面上，它的a是相对统一了，因为它是以CP为中心的，然后你的协处理器，你的加速器就最后一个interesting，你可以嵌进去，但是我们现在面临的挑战，我们硬件 a在a二这个领域s二太多了，估计目前来看还是一个挺困难的事情，所以说我觉得我们在编辑生态要形成标准，要形成一个开放的生态，可能要把这两个地方解决，解决好了以后我们可能就会坚持生生态是吧？

然后另外有一些大家讨论的就是说我们这个呃别人需要做好，可能需要把高层的一些model的信息，这些就是语义的信息怎么更好的透传到我们底下的一层层的下去，这个也是大家讨论比较多，我觉得这也是比较重要的问题。

可能我未来假定比如说去基于m二去构建我们生态的话，整个m二里面的信息的设计是吧？这个也是很关键，怎么把高层的语音信息 ahhhhh透传下去，然后里面也提到了一些相当于因为传统的编辑有很多不行是吧？Voa o 0o1020是吧？但我们现在好像 AR的编辑没这个概念是吧？要不就打开，要不就关闭，这个可能后来后面也是一个非常可以就是讨论的问题，啊嗯。我不知道有没有补充的，因为这个内容太多了我就挑了一些，我认为，然后要不我先讲完大家再补充。

然后第三个问题就是说我们提到整个深度学习的外科论文是收敛的还是发散的，如果如果因为这个还是蛮重要的，如果它是收敛的，其实我们很多策略就比较固定，我们的框架是吧，我们的编译器，我们的芯片可能就全部标准化了，但是如果它发散的事情，而且它的发散可能很发散，我们标准化就很难做。

所以说我觉得目前来看，大家我觉得主流的观点就是这些它是打算的，但是我们看到在生产场景有一些收敛，其实至少在模型的定义上把销量，但是它的前后端比如数据就会增强到后端的呃模型的应用，它可能还是比较零碎的，这个是我们的一些观点。

说话人12 03:31:00  
这句话我问一下业务上的收敛是比如说不同的这。

说话人7 03:31:10  
业务上收敛是说不同玩家都会收敛的，同样还是不同玩家，只是他在内部会收敛到。

说话人1 03:31:17  
一定的。

说话人3 03:31:19  
架构或者什么还是大家有共识的，会说给我的理解类似于比如说cmp是吧？

大家都在玩全包，那就比较好，这个事情就比较好处理了是吧？包括你说你再怎么变化，反正就这么几个基本的办公是基本上还是就那几个了，至少我们如果说在大量的生产场景，比如说 Cv LP推荐，大家有几个主流的模型，就变成比如事实标准，或者大家大量的这些东西，我觉得这个就比较好，当然我觉得可能会有新的领域是吧？

可以说新出来可能是另外一种模型，但是没关系，那vivo就是有些我固定的，至少我们的挑战就会比较少，因为我觉得你涉及的问题，他如果不care性能那就好办是吧？

你就可以这个还是玩，但是我生产产品如果是固定的话，对于做芯片的比如说编辑的就是一个福音是吧？我把这些生产产品常用的负载搞好，是不是就解决大量的性能的问题了，对，所以我下面一个问题就是m二pop是否代表了收敛。

说话人12 03:32:35  
所以。

说话人7 03:32:37  
我想知道这个答案，我做芯片的肯定是想知道这个答案。

说话人3 03:32:41  
我不知道大家有没有，我觉得我们好像没有讨论。

其实我的观点就是MFC肯定是落后于。

说话人7 03:32:48  
对最后一点，因为他肯定是比较慢，对他比较慢，他肯定是比如说他现在财政专家他或者是但肯定对我们现在没考虑这个问题，其实我觉得其实大家可以聊一下，就是说它慢，比如说是它也在变对吧？我们是有几个没用，对吧？它可我们可能认为它有几个 model没用了，但是有可能再加一两个是不是就基本有代表性了？我不知道是不是大家如果说收殓我理解，那也不会太多。

说话人3 03:33:21  
对，我觉得反正这次应该两个不同的问题，我觉得收敛那个问题可能大家看看 bug到底是不是趋同，对吧？未来至少从趋势看，然后mrf我觉得这里面有很多玩家的例子，本来就是说安排客户本来就是有一帮人是吧去定这个标准是吧？

然后来他肯定会根据他的利益去制定两个，我觉得对。

说话人7 03:33:53  
其实我不他那个东西我也不认为它就是有代表性，我这上需要代表性的是什么。

说话人3 03:34:02  
你也可以说他都不对，我觉得至少你比如说现在就NLP你用枪的方法这个事情，大家还是基本上就实际就是这么对对对，反正他对吧没人会跳出来是吧？

然后就已经倒也没有，只是我觉得是我们下面的就是一种讨论而已，o。

说话人12 03:34:26  
好谢谢，因为风险的观点。

说话人5 03:34:30  
这个问题其实我们这边有过一些类似的讨论，其实应该是因为发起这个事，问题就是说可能我们有些boss非常flight的时候，可能就像说一个检测模型，可能30%是在投资百分之外，可能在一些非常弗莱克算子里面去了。

说话人3 03:34:47  
那么对于这样的算子的话，可能大概率是不太容易直接被加速加速度。

说话人5 03:34:53  
如果为每个算作我们回到CPU里面算的话，计划svo现在我们可能拍套式里面的话，我们就说一个CEA的programs启动起来，在里面一些分析浪费出来的科目，那么这里面我们有没有一种可能说我们对这些我们能够看到的比较重要的，我们对背后的算子的在一个更细的粒度的第一个规约，这个规约可能不是在这种可能拍到ok的力度，而是在一个类似于像hr这样的一个类似这样的r的里面会约，如果发现哪些算子，在这里面描述不是很方便的话，我们来进行扩充，最终是能够形成一个封闭。

如果封闭机存在的话，其实我们对任何一家新建公司，我们把结合支持好。

说话人3 03:35:38  
我们也讨论过，就说不行，全讨论一下，下面会讲一下，但是我觉得你这个问题可能里面有两块，一块就是说。你是属于前处理后处理，还是属于搬个棒的对吧？这个可能还是要分开对吧？另外一个就是说这个不规则算是你筛选用的还是生产产品用的，这可能也要分开都要加工。

说话人19 03:36:08  
其实我觉得这个应该好几个组应该都讨论过，包括最前面提的一个问题，有没有存在一个算式以下，然后三级以下的一个叫perfect的东西，能够把所有的都拼起来，我理解其实如果要真的要全，他就需要是独立完备的。

说话人15 03:36:28  
我们讨论过，但fooling fae我可能我表示反对，因为我觉得其实 different model还是比较restricted的，对他不一定真的要处理吗？

说话人19 03:36:39  
对我其实这样，如果说你想要把这个问题完全的close掉，图灵完备你就close咱们就不用纠结了，对，但实际上要做的话，我们肯定没有那么多成本能给他做到重新分配。

事实上可行的是说我们在一个受限的范围内，让它能够表示这个就是99%的这些case，这个应该是可行的，但看怎么去，我觉得我们是不是可以去共同努力去造这样一个东西，就是一个permit，它能够表示目前特色flow或者排他期里面99%的case。

然后有了这个case之后，这一层其实它就可以把硬件和上层的框架一定程度上做一个划分，然后大家可以交标准化的产品去做工作。

说话人3 03:37:32  
现在反正我记得mi本身有个项目是那个答案，我们担当对，还有反正尼克斯本来就是做这个工作了，不是，所以说我为啥说其实很简单，那个是成功了现在。

所以说你觉得如果他成功了，那就搞定了，如果比他没成功，那就是他有问题。

说话人7 03:37:58  
到时候肯定有点这些好的问题，他有可能用那种比较抽象一点的，比如说data里面对蒜子做了几类一对一的，多对一的一对多的，是吧？

然后他其实。

说话人1 03:38:12  
用一个所谓的重索引语音的词来描述这种，他不是给一个具体的算的是什么形式，而是说明代词大概是一个好像是最多的了忘了，但是你可以在里面再写个小的拉姆基础的方式，然后一算的话它种类很有限是吧？

说话人7 03:38:37  
我是觉得其实本来我们发起论坛或者发起讨论的这是一个话题，我们能不能通过反正但是你说往哪个成功不成功，或者他为什么成功不成功，我们其实不太关注，我们是想说通过大家讨论我们是不是能讨论出这么一个东西，啊如果能讨论出一个或者大家基本上是有一定共识的，我觉得就有一定代表性，对吧？

如果当时讨论我们这些人都讨论不出来一个共识，可能确实也。

说话人15 03:39:05  
对吧？

我们其实还是很有代表性，大家比如像安徽像雪峰，你就是安防的成功，这个问题还挺有意思的，其实。

说话人5 03:39:13  
我自己印象汪意思在深度学习框架没有出现两家不大的情况，肯定是很抛开的，问题，大家都希望去介入它，去用它来做一些中转和一些联盟的工作，基本上进入到PAD，已经超过了这个服务以后的话，至少Facebook内部是不再去推这个东西，因为它已经没有incentive了。

所以这个时候的话它出于各种各样的一个发展，我的套吹不起来。

因为出现了说其实现在网点最大的贡献者应该是Facebook之外，如果我没记错的话，应该是以微软为主来贡献更多一些现状，所以其实确实如果说作为一个中立标准的话，它不能算很成功，那佛洛也不是那么待见他。

说话人3 03:39:49  
拍照也不是那么待见他。

说话人5 03:39:51  
但是有趣的现象，几乎所有的硬件公司都会先做二次对接，包括推理包括训练，因为大家发现它太neutral了，非常适合我们的戏，尤其是特别好玩的一个问题，对，所以这个问题也就是说它不光是一个技术问题，还是一个商业问题。

举个例子，如果说我们想干这个事情的话，假定说是我们拉着，因为像通报的话属于国外的公司，如果我们拿着money ball来去聊的话，可能也在想像money ball的角度会有这样的动机去干这种事情吗？

我的感觉是有可能东西不会太强，因为一个框架好像没有太大必要说我花太多巨大的精力去打通跟别人对接，而我是应该不断扩大我自己的市场份额，所以这个确实是不是一个纯技术的 problem给我的感觉。

说话人3 03:40:38  
对杨军说的反正我觉得是一方面，另外一方面我们也讨论过，就是说其实跟深度学习产业相关，因为我们实际上我也参与过很多标准的讨论，比如说我们做4G3G的标准是因为大家坐下来谈是吧？讨论完以后把把大家写定说好，然后我们干5年对吧？然后下一个5年来下一代再谈，但是深度学习不是这种玩的模式，就是说我明天就做多了一个op然后你加不加？

你来不及加是吧？来不及加怎么办？

我就自己搞，自己搞的到最后就发现与其这样，我为啥要这样做，所以说这个是一个非常现实的问题。

说话人12 03:41:28  
当然如果我们。

说话人3 03:41:30  
所以这是个挑战，我觉得不是说不行，我们今天主要是把这种挑战说出来，我觉得我们。

说话人7 03:41:38  
就不谈标准，其实我们没必要，我们这些人其实很开售的讨论，我们不谈标准，我们也不想推标准，我们只是想讨论出来有没有一个合理的下一个问题对吧？

对，然后我刚才想说的是说这个东西太快了。

说话人3 03:41:53  
我可能今天我们讨论完可能明天又出来一个，这时候大家这个东西要不要遵守，然后如果不遵守大家可能自己参数自己弄完，然后结果是两家不一样，然后又放在一起又来安排一下，这是个线，这是个问题。

说话人7 03:42:09  
不是说我们愿不愿意做，就是说我们不是把它作为标准就是说我们不存在遵守不遵守的问题，我们是一个recommendation，我们是大家这些做起来调有做硬件的，有做软件的，大家认为反正在现阶段这个基本刊物了多少，就形成一个反正我们讨论出来的一个东西，也没说谁要遵守学校，甚至我们大家自己知道对吧？

因为其实我们从现在公司我们希望得到一个稍微对吧稳定一点的答案，你后面加可能也没关系。

说话人12 03:42:44  
但可能比如是。

说话人7 03:42:46  
是隔一年我们再更新一下，这个没问题，我就是纯技术讨论，我们也不想说是推什么标准，谁来那个没意义。

说话人13 03:42:54  
对，其实我觉得这个我比较赞同杨军的一个观点，就是这个事情我觉得他可能不是光技术的问题，它是一个商业的问题，而且我觉得说确实我们也没有必要就是说一定去形成一个比如说一个算子的具体一个例子的是吧？

我觉得那个东西说实话说一一时半会也形成不了，但是其实我觉得还是我一开始说的关于我觉得我们可以做的事情就是说我有没有办法可以要加快收敛的过程，我觉得可能是一个比较我们可以去想这个事情是吧？

确实我我短期内我是肯定找不出这么一个历史，但是我有可能可以做一些事，可以去加快历史形成这个过程，也许最终形成的例子不是由咱们来定义的，有可能是比如说某某一家大公司最后形成事实标准，它出了一个例子，但是我们可能就是说可以在这个过程中有可能可以去出一些利益，或者是做一些事情能够促成这个过程，我觉得这个可能是我们可以去做的一个事情，对。

说话人7 03:44:00  
我看一下为什么在短期内找不到这个意思。

我说一下我这样的还有两个硬件业务，我其实这样。

说话人19 03:44:09  
如果说要找一个训练的就是要用到的蒜子，一个列表的话其实是存在的，你把排气或者跟弗洛的API表达下来，他就是一个事实标准。

对，但这里面的问题是说，不管是拍大姐还是特斯拉，他们的算子现在都是基于MV的扩大生态来编来编程的。我们要的前面提到的，我在AP的基础之下给出一个新的permit Abel，然后所有算子这1000多个蒜子都能基于这些p来表示。

我我我观察到是说这里面其实是一个技术人力成本的问题，我首先假设存在这样一个 QQ我设计出来了，我也需要把这1000多个算字在 premium上全部重新实现一遍，这个里面的人力是没有任何一个独立的实体能够去commit的，你想拍那几部分的话都是社区这么多年这么多人往里面去贡献的，代码怎样还参差不齐。

说话人12 03:45:10  
是。

说话人15 03:45:15  
我其实能follow一下深圳的建议，就是说有没有可能能够简化这个流程，让这件事情变得可能，对吧？我提一个例子就是迪欧派克，因为迪奥派克是天奇对吧？他发起的说我要让所有的框架之间的cancer的in memory能够写，你可以讲 Iphone也不小，right？但他为什么能够work，但至少目前很多人还是用的，我觉得有两点，第一点是它非常简单，我要加迪奥派克，我只要加两行两个函数就结束了，这件事情就这样，cost非常小，这是第一。

第二就是说它有未来它很多议题是把它放到后面去decid，比如说调派了现在开始去讨论说怎么支持simple as memory strain等等这些原来不靠谱的东西，就说他一开始的 slope特别的小，这件事情就是让整个社区，我觉得这个东西当然他也是拉了很多当时的跟他方向的创始人过来一起来到什么对吧？

但是大家觉得这个东西iphone的还好，可能花一个下午我也搞掉了，right，所以他就到然后到流血滚起来了，因为比如说我们现在去看要看生态里边像cvor ele就很支持他很多个PK全部是jdk，这样它能够很方便的去place一些可能，然后生态起来以后，它反过来去迫使这些框架开发者他必须去维护这个东西，没有办法，然后我再去说我在这上面再加新的功能，大家也更好接受。

说话人2 03:46:45  
一些。

说话人1 03:46:49  
所以这个有点像一开始给你免费使用是吧？成本很基本要钱以上，年薪很高之后你的不代价。要不我继续其实还有一个我们讨论了一下，就是说。

说话人3 03:47:09  
因为我们现在讨论的更多的是说编辑或者说是op领域的接口标准，有没有可能在边框架的。

说话人2 03:47:19  
接口。

说话人3 03:47:21  
做一个标准是吧？如果那个标准说那不就更一劳永逸了是吧？向下就更加通顺了，当然更难是吧？

当然我们也讲了一些路径，就是说比如说你南派加这个为的是的是一个最简洁的情况，因为我觉得第二个事就是这么做的消息啊，然后大家所有的其他的库就基于这个东西去往上扔，其实也是一个建议，当然我们大部分的人觉得包括敏捷欣然，包括来双方都认为就是说现有的框架的遗留系统的带来的问题比较多，或者说他这个包袱比较大，这个事情还是挺难推的。

现在实际上我们这个框架的结构基本上都是配套起来那一块是吧？就有点像讨论，但又不一样，绝对是这种，所以我觉得这也是一个一个问题。然后下面刚才讲的就是嗯在这个模型标准上，我们现在可能有两个大家认为是比较公共的，一个叫Alex，第二个我理解m二也在推这个标准是吧？

模型层面的标准这一块我觉得反正我觉得如果能够统一是个好事情，其实因为我们最早马克思抱怨的是奥尼斯最早，但是我们确实发现我刚才说的那个问题，我刚才不是说反对这个事情，而是说我说这是个挑战非常大的挑战，因为大家奥尼克斯变动太多了，最后奥尼克斯不是大家想象的奥秘，就是你跟我讲，我给你一个，但是实际上你给的跟他给的是不一样的，这就是一个问题。对，我觉得后面确实如果说我们要定义模型的标准，是可以需要一方面技术上去讨论，第二个可能机制上也要讨论。因为现在像aka，社区要提交一个更改，那是很漫长的事情是吧？大家要这个社区要讨论，这实际上不太认可。

然后最后一个现在提的训练框架，是一个未来跟动态还是跟静态是吧？然后有几派观点，可能从新来的角度来讲，可能动态图从灵活性包括效率上会更高一点。

然后我当时提的一个观点是说可能要考虑两个，一个主要是我们这个硬件厂商的生存空间的问题，是吧如果真的全是动态图的话，其实dsa架构就发挥不出它的效率了，所以说我觉得这个问题可能最好分场景考虑，就是说我们一般可能可以把一个AR的这个过程分成研究阶段跟生产，区的研究阶段，你可能需要的是说更加的灵活，效率更高，你的模型变化的更快更多。

另外一个就是生产阶段，生产阶段就是说你是要大规模生产，比如说像互联网厂商你的一些实时的这种推荐是吧？推荐搜索，包括你的大量的这种语音的这种语音语义分析的这种场景，你可能是搭了一个几万台的这种集群在跑，这时候你肯定对能效比包括性能是要求很高的，然后那个模型假定我们生产阶段的模型是能够收敛的，这两个产品就可以分开解决了是吧？

我在研究阶段可以用你的动态图性能低一点也没关系，然后你生产阶段可能等你研究阶段的模型基本上搞定了是吧？到生产阶段部署的时候，你可以用性能更加高的动态。

同时就是说因为现在动态图主要是在铝板上传，其实说白了未来GPU到了100甚至更厚的这样探测call以后，其实它的性能调整也会非常大。

说话人1 03:51:38  
好结束了，好谢谢，这个地方我补充一点啊我的一个想法，就说。

说话人7 03:51:46  
这里边提到这个意见这块发生欢迎静态对吧？

其实对于我们来说，我们之前有个实践的，后来我们也是逐渐的认知。

说话人15 03:52:02  
就说我们现在加速器这样布移做的过大，就是说你就在你加速器做本身是很大的，但是你里边的扩规不一样，比例扩大，我们一般加速器的话它都会有一些。

说话人7 03:52:17  
定型度，因为我们处理的case可能有的是做三维的，有的做二维的。

说话人15 03:52:23  
对吧？

有的人也有意味着向量预算，实际上就说美国维度我们可能如果说我是三维的三维的，我处理的数据本身可能每一个维度都不一定正好把它抽回来，比如说每个维度都用了90%，两个维度90%，一层就成了81%，再一层就70%多了，利用率也降得很快，所以我们就说第一。

说话人7 03:52:49  
加加速器我们尽量的同样的我计算能力，我尽量拆成很小很多小的框，然后我线路大一点的线路上我就可能用10个扩，小的线不能可能用3个扩，因为我就是把扣一个填满，最后可能要做填满，基本上这样的话我们就觉得就是说从硬件的计算效果来说，它多多少少能够把胆子没有泄露这个问题能够做得更友好一点，给你提供一些感谢。

说话人1 03:53:22  
那么我们由于时间关系，由于时间关系，我们这个也到了尾声，今天其实但是还是有微信，还有很多问题还可以值得继续讨论，那么这也是我们后面要去搞开源的，开始sorry是搞开放式的，这种研讨会上主要的动机之一，那么我们后面可能关于研讨会怎么去组织，我们可能更多还是在线下，我们可以在群里面稍微讨论一下就行了，没有必要今天把他留在这里，非得把讨论的一清二楚，那么我们今天这个论坛就到这里结束了，然后我们最后就合影，然后大家纪念一下我们今天这次 QQ的一个比较有标志性的事件，对吧？

有可能将来会产生一个非常大的影响。

说话人7 03:54:15  
请各位嘉宾我们嘉宾广播到台前，包括群众朋友们也一起来合影。

说话人31 03:54:22  
这个是大家发言拿药好，然后就感觉我说的小时候还上不了，可能还得做一下。

说话人7 03:54:33  
可以在班班里是吧？我觉得这肯定是那么快。等一下上来把他们做了，前面坐一排，后面站一排，然后台阶上再站一排，先这样不见了，那就这样坏了，要了过去还比较像传统的学术会议，我们就下面干一个上面干一个，然后昨天大家自己能两200往前走。

说话人31 03:55:05  
可以大家知道你们中间我们了可以吃饭还是回去？

真是回去了，我都我们我们。

说话人1 03:55:13  
插一下，我们拿起一个我们是主办方这个事今天没有。

说话人31 03:55:21  
把情况出来。

说话人7 03:55:26  
可能要在大数据，但是还得往后，好可以这边也过来，我们这边也大家就是说你不知道才上，然后自己查控把两个出来，前面可以分两个人，刚开始那边。

说话人17 03:55:40  
可以来两个空间，这边可以后面如果你看不到的就在前面几个。

说话人31 03:55:46  
对。

说话人7 03:55:47  
没有两两两位感谢来看着对吗？还要再往后走吗？

说话人17 03:55:53  
可以在前面再登记我就可以了，那过来我们等一下别拖时间，别这样，什么都可以的。对没事这样对边上起，好来两位把这两位好在这边。

好这样后面还是有点看不到的，你们如果觉得看看到看不到前面的话，你们就看一下。

说话人7 03:56:12  
看下来你们到前面对到前面你说，对然后后面稍微再慢一点，复刻一点。

说话人15 03:56:21  
所以你用手机帮我拍一下，ok来看这里。

说话人17 03:56:32  
好，被后面脚稍微垫高一点，好吧，再往前来一点往前来，好，可以来这里了。好，可以来。这一张很好。好一些点赞来好吧，点赞来着，好戴帽子的帅哥稍微再定高一点，好可以来分别了，所以好好可以，好谢谢。

说话人31 03:57:00  
好的，我再继续发一下那。这个地方说一下，你看d的固定的知道了。