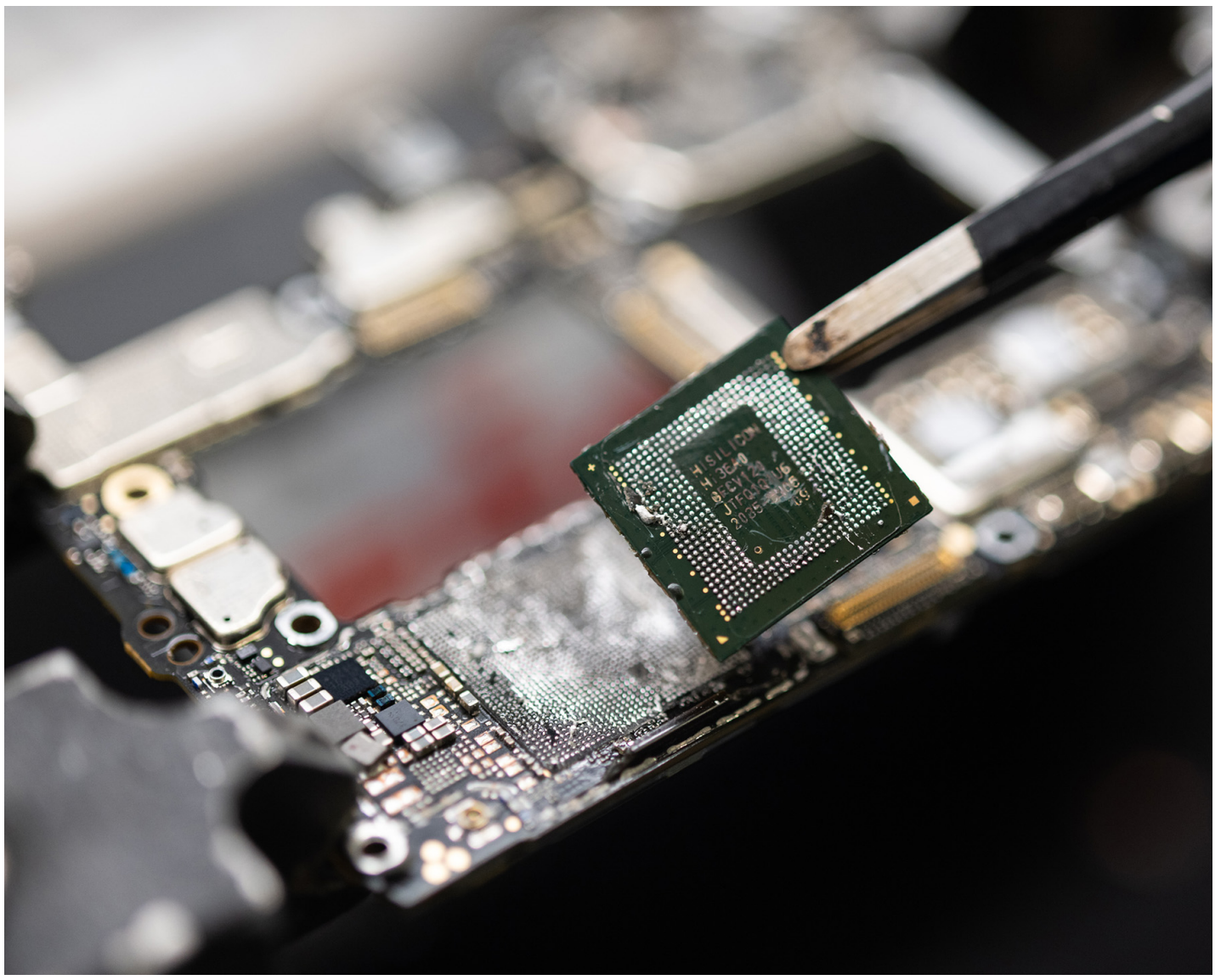
评论国际大陆科技深度

在“算力即国力”的今天，美国最新的技术封锁给中国带来怎样的压力？

美国选择在“习拜会”之前发布这个重磅炸弹，完全不考虑中国可能的反弹，就是释放出“别的都可以谈，芯片免谈”的信号。



华为Mate 60 Pro手机芯片由中国制造，以突破美国的芯片管制。摄：James Park/Bloomberg via Getty Images

杨路

刊登于 2023-11-14

#芯片 #出口管制 #半导体 #中美关系 #中美科技竞争 #评论

分享全文 0

今年10月16日公布的美国先进计算出口管制更新规则（全称是Advanced Computing/Supercomputing Interim Final Rule，AC/S IFR），以及相应的半导体制造设备出口管制更新规则（Semiconductor Manufacturing Equipment Interim Final Rule, SME IFR），绝对是中美关系史上，乃至当代科技史上的一个“重要里程碑”。

与2022年10月7日发布的上一个版本相比，这一版本的技术封锁在设计上严密了许多。虽然英伟达（台译：辉达）仍然找到了一丝缝隙，又推出了H20这个新的中国特供版，但美国政府已经用实际行动宣示了，未来的管制红线，会频繁地移动，企业可以操作的空间只会越来越小。这意味着中国用户能够买到的英伟达芯片总体性能未来只会越来越差。

在算力战场，中国未来只有自主研发一条路可走，这当然不一定能成功，但一定会产生大量的赢家和输家。

大模型算力之争：我们这个时代的核军备竞赛

如果“军用级”人工智能（虽然具体形态当前仍然模糊）是核武器，那么算力芯片就相当于浓缩铀。

以ChatGPT问世为分水岭，未来的人工智能发展将分为两种道路：有算力的（the GPU rich），和没有算力的（the GPU poor）。

如果你有足够算力，比如一万张英伟达H100（2023年发布，台积电4纳米工艺，单卡算力900 TFLOPS）AI加速卡，那么你可以用3个月的时间训练出一个类似于Open AI GPT4（1.8万亿参数）的基础大模型；反之，如果你没有足够的算力，比如你手上只有一万张英伟达 V100（2017年发布，台积电12纳米工艺，单卡算力130 TFLOPS），那么你就需要2年的时间才能“炼”出一个GPT4——基础大模型正在以差不多半年一代的速度疯狂发展，两年训练一个大模型在现实应用中几乎没有任何价值。所以你只能用别人的大模型，去搞一些细分赛道和应用领域的“小模型”。对于企业来说，小模型也可以有自己的商业模式。但在国家层面，特别是以当前中国的战略取向来说，无法研发“自主可控”的人工智能大模型几乎是无法接受的。

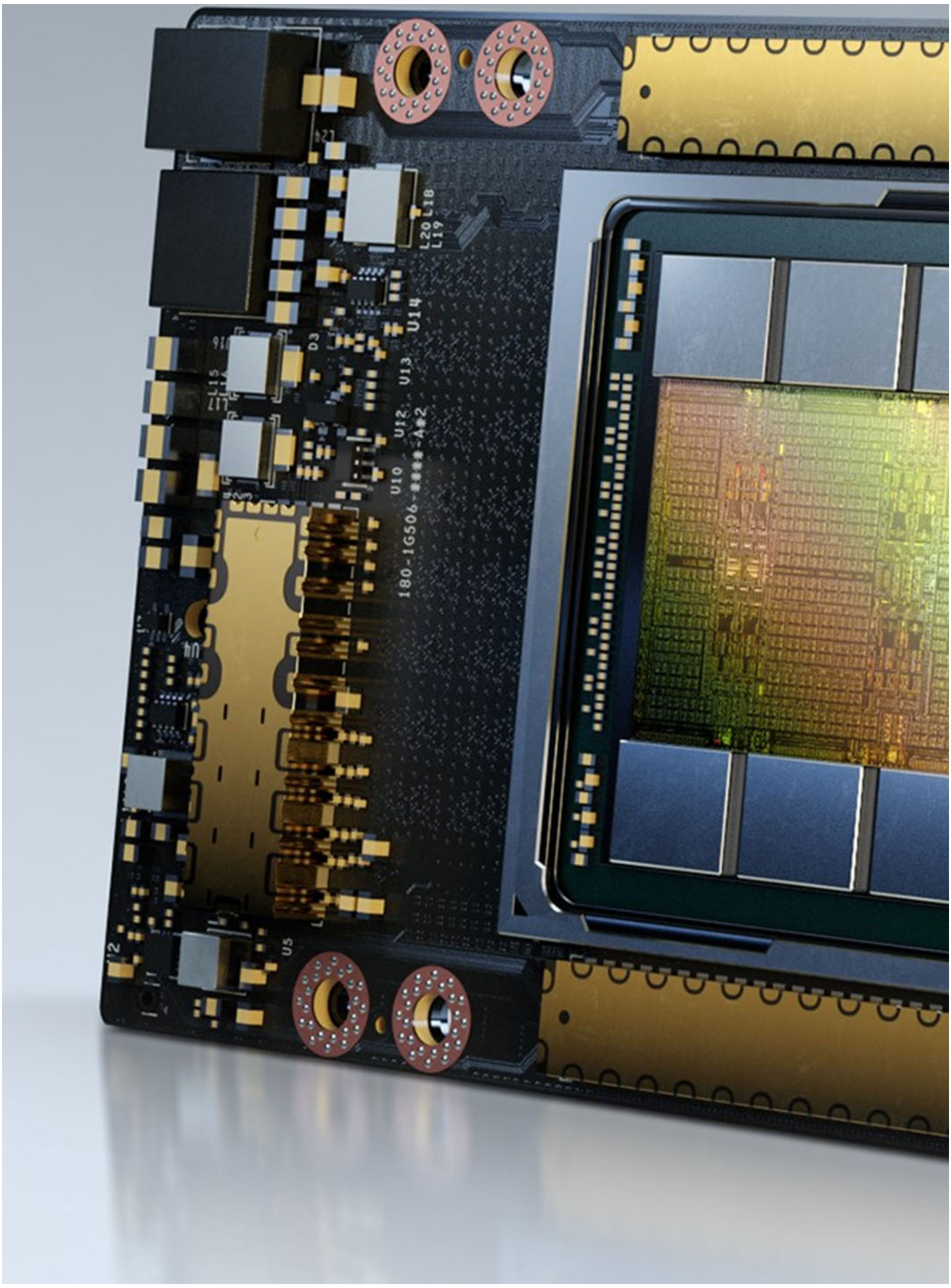
用地缘政治视角，人工智能大模型竞争就是我们这个时代的核竞赛。如果“军用级”人工智能（虽然具体形态当前仍然模糊）是核武器，那么算力芯片就相当于浓缩铀，而半导体制造设备则相当于离心机。

A800“漏洞”

“中国”和“Big Tech”在今天美国公众舆论中几乎是两匹死马，人人皆想踢上一脚。

在2022年10月第一版先进计算出口管制规则中，美国政府将先进计算芯片的红线设定在“单卡算力超过4800 TOPS，且双向传输速率超过600 GB/s”。“4800 TOPS + 600GB/s”这一指标显然是来自当时市场上最先进的英伟达A100加速卡。我记得当时政策出台之后，各路分析师的一个争论就是这条性能红线未来会进一步收紧吗？我当时的看法是不会。

从结果来看，我显然是判断错误了。在2022年新规发布之后，英伟达非常“机智”地推出了算力与A100相同，但传输速率略低（400GB/s）的A800芯片。A800是一款“中国特供”芯片，这是一款纯粹为了绕过出口管制而存在的产品。其设计性能只是刚刚低于管制红线，因而得以合法供应中国市场，并成为中国用户能够大量买到的最优性能产品。英伟达也通过这个办法，尽可能补偿了A100无法销往中国的收入损失。



NVIDIA A800 显示卡。图：NVIDIA官网

放在一个其他时空下，英伟达这样的操作恐怕不会引发太大关注，毕竟这个行为本身合法合规。这样的做法在跨国企业中也很有代表性，在法律红线之内追求最大的利润空间在全球化高歌猛进的年代里再正常不过。

但今天不同。英伟达的算力卡在今天几乎是“算力”的代名词，是人工智能开发者眼中的“硬通货”。A800虽然传输速率低于A100，但其中的带宽差异在某些情况下并不会明显影响工作结果。更重要的是，“中国”和“Big Tech”在今天美国公众舆论中几乎是两匹死马，人人皆想踢上一脚。英伟达作为一家全球业务四分之一来自中国的大型科技巨头，自然又是舆论重点关注的对象。A800这样的“小聪明”在这样的语境下就引来了大量美国舆论的挞伐。对于许多意见领袖来说，A800就是美国出口管制漏洞的写照。而这个漏洞，必须要堵上。

新指标的引入：出口管制交叉火力

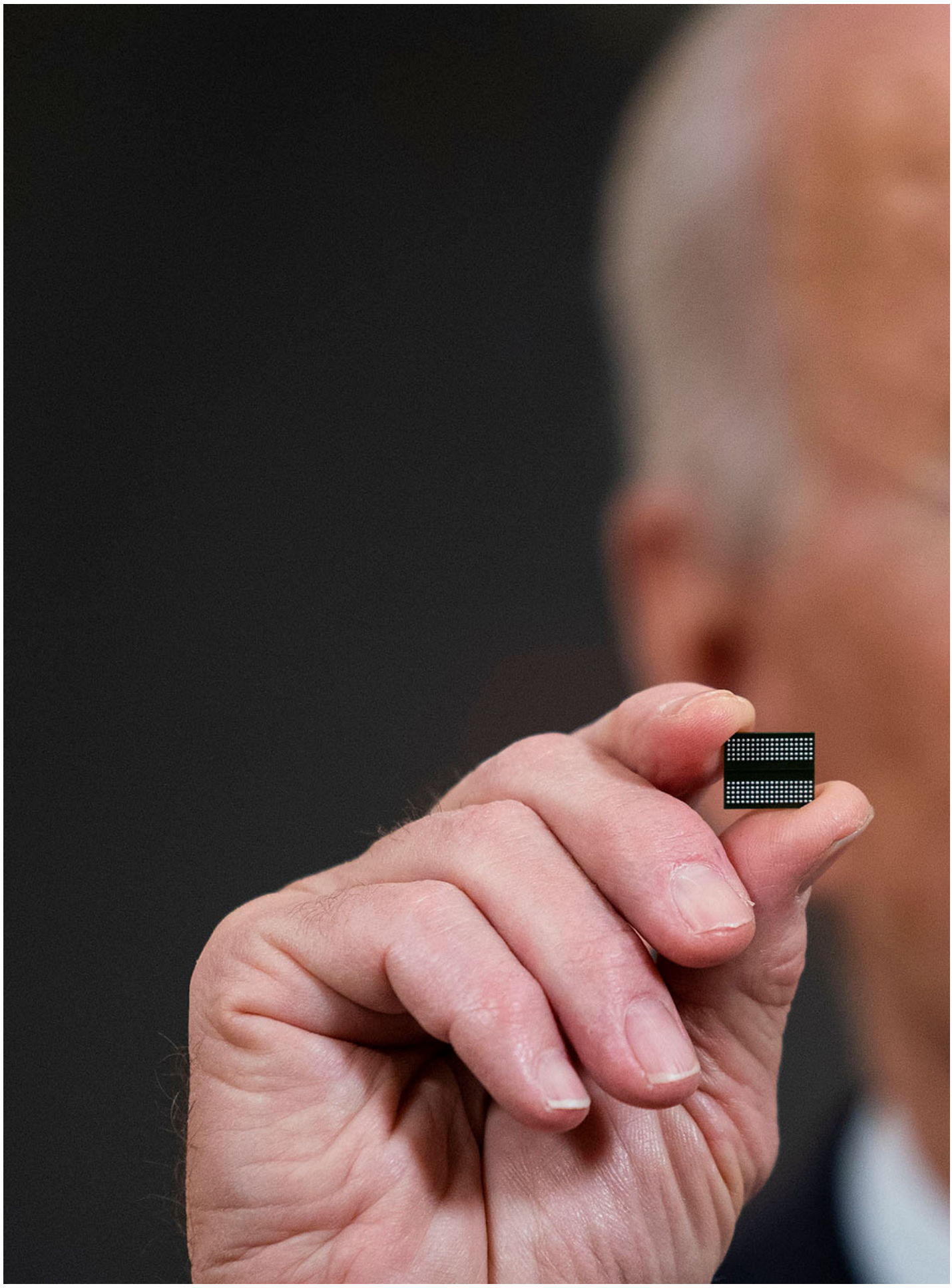
可以想象这次新规背后的政策团队，在这一年时间里，花了很大功夫做研究，所调动的资源也超过了美国非军情部门常见的标准。

在此次规则更新中，美国商务部通过一个相对技术性的手段来解决了A800问题。简单来说就是，它将此前的“算力4800 TOPS且传输速率600GB每秒”这两个需要同时成立的条件，换成了一个单一的“总处理性能”（Total Processing Performance，TPP）指标。

TPP即算力与位宽的乘积。举例来说，如果一款芯片的设计性能是“INT8精度下算力600 TOPS”，则其TPP = 600 x 8 = 4800。与上一版本不同，这一指标只与算力有关，而与传输速率无关。也就是说，A800和A100的TPP是完全一样的。这也就达到了堵住A800漏洞的目的。当然，取消传输速率指标也为英伟达寻求新的中国特供产品空间留下了伏笔，这一点下面会讨论。

本次出口管制红线调整的另一特色是引入一个新的指标：算力密度（Performance Density, PD)。算力密度的定义是一颗芯片单位物理面积所对应的算力。以英伟达A100芯片为例，其算力为4989 TPP，裸片（die）面积为826平方毫米，对应的算力密度即为4989 / 826 = 6.04。

对于电子芯片来说，决定算力密度的最大变量来自于单位面积上的晶体管数量，简单来说就是制程。因此引入算力密度作为红线，实际上是在管控制程技术。此次规则按照“总算力—算力密度”组合的方式设置了三档阈值（5.92、3.2、1.6），大体上对应的是7纳米、12纳米以及28纳米制程。这与美国在政策规划中将28纳米视为“先进制程”分水岭的做法是一致的，确保了不同规则之间的技术一致性。



2021年2月24日，美国华盛顿，总统拜登手拿著半导体晶片。摄：Doug Mills/Pool/Getty Images

引入算力密度指标的另一个政策意图很有可能是封堵中国利用芯粒（chiplet）技术方案绕过出口管制建设大规模算力集群的可能性。在芯粒技术中，开发者可能利用die-to-die封装方式，提高每个芯片能够承载的处理器单元数量，从而提高实际算力。壁仞科技（此次同时被加入实体清单制裁）2022年8月发布的BR100芯片，即采用了芯粒技术达到了高算力。芯粒技术此前一直在国内被寄予厚望，被很多人视为突破“英伟达—台积电”技术霸权的一个机会。而此次的规则改变，将大大提高中国企业继续开发芯粒产品的难度。

在算力和制程出口管制两重“交叉火力”限制之下，此前市场上所有主流的人工智能算力芯片都将成为出口管制对象。在老产品中，红线之外的只有英伟达V100和谷歌TPUv3，两款芯片分别发布于2017年和2018年。如前所述，大模型几乎半年更新一代的速度面前，用五年前的芯片训练最新的大模型耗时将以年计，几乎没有实际价值。

与去年发布的“1.0版本”相比，这次的先进计算出口管制“2.0版本”在技术细节上的考量和设计，有极大的提升。可见背后的政策团队，在这一年时间里，花了很大功夫做研究，所调动的资源也超过了美国非军情部门常见的标准。需知，2022年第一版芯片出口管制出台的时候，ChatGPT还没有发布。过去这12个月，全世界的人工智能研究，正在以史上最快的速度演进。而今天这个版本的出口管制在技术层面基本上反映了最新的技术进展。对于一个政府官僚机构来说，相当惊人。而美国选择在“习拜会”之前发布这个重磅炸弹，完全不考虑中国可能的反弹，就是释放出“别的都可以谈，芯片免谈”的信号。

英伟达和美国政府的打地鼠游戏

当前的“芯片战争”的大背景自然是中美博弈。但在这个大故事下面还有一层英伟达，以及其他类似情况的跨国企业和美国政府之间的博弈。

然而英伟达还没有放弃。正当大部分人——包括我在内，以为新规出台之后英伟达的中国算力卡业务即将全面断绝的时候，英伟达展现了极强的供应链管理能力，在新规发布一个月不到的时间内（去年推出A800也是在新规之后极短的时间内，应该是有备而来，可见英伟达政府关系部门的预算也不是完全没有效果）推出了一款新产品：H20。

H20基于英伟达的H200卡——也就是旗舰级别的数据中心芯片H100的同系列产品进行了完全针对出口管制的修改。简单来说就是，虽然新的出口管制规则限制了总算力（TPP）和算力密度，但是放过了传输速率这个指标。而今天的人工智能大模型训练，和一张GPU就可以居家操作的比特币挖矿不一样，动辄需要上千卡规模的算力集群，最终的训练速度往往取决于硬件性能“木桶”中最短的那一块。这个短板在一些情况下可能是算力，在另一些情况下可能是其他条件：比如存储、带宽，或者算法。

总算力和算力密度只是这个“木桶”众多组成材料中的两块木板，现在确实受制于出口管制无法继续加长，但这并不是唯一的两块木板。在今天的很多实际应用中，另一常见的短板就是传输带宽。大型的算力集群往往受制于卡与卡之间的数据传输通量，而无法发挥单卡最的算力效能。

H20的峰值算力是2368 TPP，对应的出口管制算力密度上限是3.2。而这款产品目前公布的算力密度是2.9，因此如果按照法规字面解读，是不受出口管制限制可以合法销售给中国客户的（美国商务部还没有回应）。H20大幅提高了存储容量和传输带宽。虽然其算力远不如H100,但其存储容量超过了H100的80GB达到96GB，传输速率通过英伟达独有的NVlink技术达到了900GB/s，这就与H100相同。

整体来说，H20的产品策略是通过提高传输速率性能短板，弥补算力不足的缺陷。在推理场景下，H20可能比H100效果更好。但在政府层面更关注的训练场景下，H20的性能比H100肯定还是要差一些（目前定价不明，所以实际性价比很难说）。但绝对性能在这里不是最重要的事情，新规发布之后H20仍然是中国用户能够合法并且成批量买到的最佳产品。其意义类似于去年的A800和H800—大家都知道这是“阉割”版，但仍然是最优选择。

2023年5月29日，台湾，一名男子拿著NVIDIA执行长黄仁勋签名的显示卡。摄：Ann Wang/Reuters/达志影像

这一情节转折令人惊叹。英伟达展现了无与伦比的“见缝插针”能力，几乎从规则的一道小小缝隙中开出了一辆卡车。并且在这么短的时间内重新组合了供应链，准确拿捏了中国客户的需求，推出了几乎无法拒绝的产品，其产品和市场能力也完全配得上这家“世一芯”企业的声誉。这也再次说明了，用政策工具来管控前沿技术是多么的困难。

当前的“芯片战争”的大背景自然是中美博弈。但在这个大故事下面还有一层英伟达，以及其他类似情况的跨国企业和美国政府之间的博弈。目前看来英伟达们虽然明显处在下风，毕竟个人奋斗在时代进程面前往往无力，但是英伟达们很有可能会坚持到最后一个回合，这个拉锯可能还会持续很长时间。

H20和A800类似，本质上是一个权宜之计。可以预见的是，在下一个版本的出口管制更新中——不要忘了，美国商务部长雷蒙多此前可是说过 “每年至少都会更新一次”，很有可能出现针对带宽的限制，来堵上H20这个新漏洞。但在那一天之前，可能是六个月，可能是半年，英伟达一定会尽量卖，中国用户也会大量的买。因为双方都知道这个生意的保质期不太长，能买（卖）多少是多少。

都在和时间赛跑，中国买家等的是国产芯片什么时候能够赶上来，在那之前一定还是尽力采购英伟达，确保现有的需求不断档；而英伟达，一方面是能做一天生意就做一天（下一个季度的营收目标现在看来是保住了），另一方面肯定现在也在琢磨下一个可以利用的“短板”在哪里。

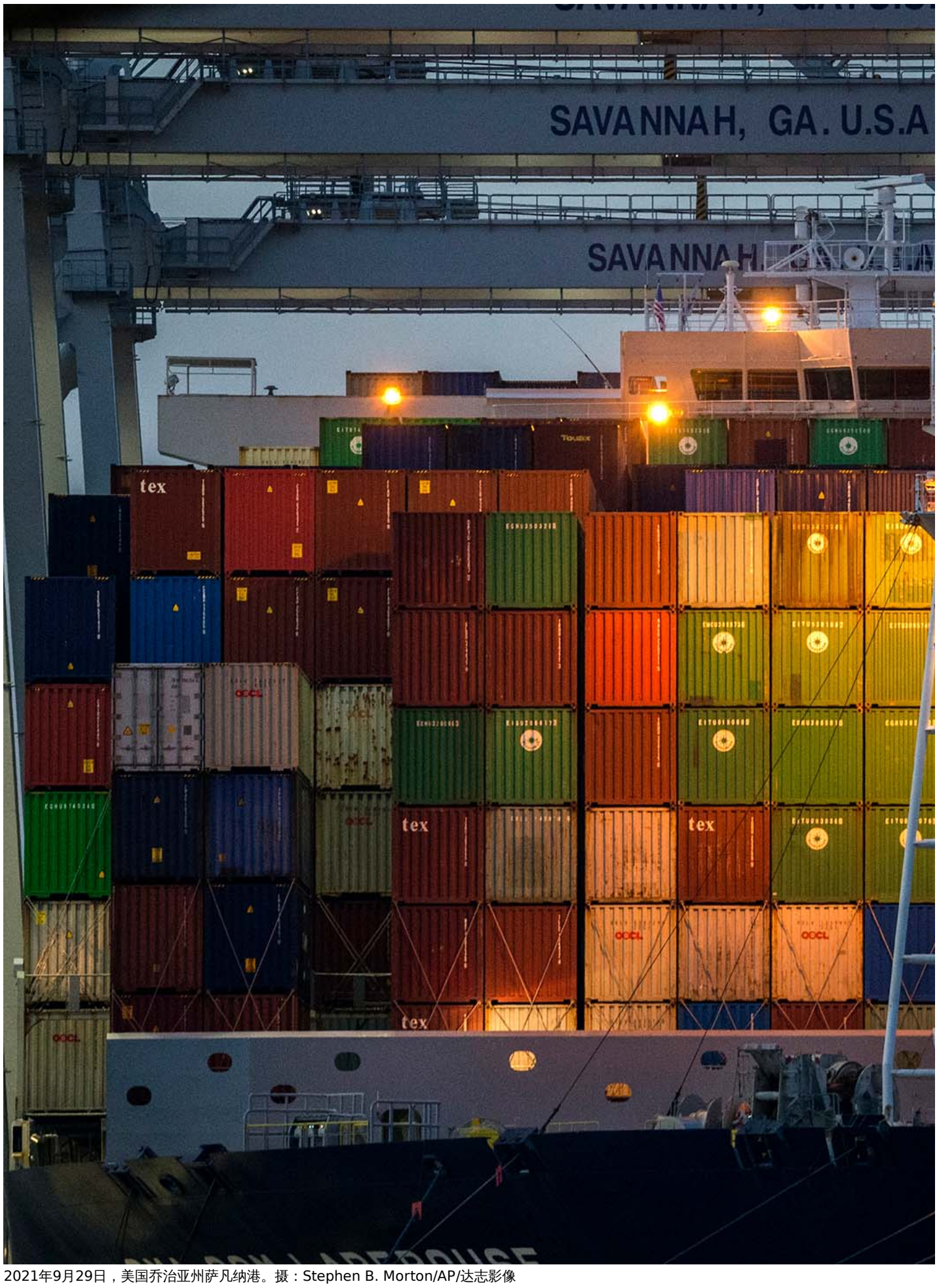
赢家和输家

> 算力即国力，谁手上有卡谁发财。

基本上美国每更新一次科技制裁，中国的相关行业就会产生新的赢家和输家。

输家自然很容易找，这次“上榜”（被实体清单制裁）的一家企业最近几日已经宣布裁员。被实体清单直接制裁不仅仅意味着不能够获得美国技术，更会影响声誉，失去供应商和合作伙伴，提高融资成本。

而对于那些在2022年规则之后，战略上下注A100性能标准红线不会移动，继而推出了性能无限接近红线的国产GPU公司来说。红线此次实质上收紧，就把最近这12月内问世的许多款采用台积电工艺的12纳米芯片（有不少）摆到一个十分尴尬的位置。



2021年9月29日，美国乔治亚州萨凡纳港。摄：Stephen B. Morton/AP/达志影像

对于高性能芯片，美国采取的是“外国直接产品规则”（Foreign Direct Product, FDP）。也就是说，即使这些芯片完全是由中国公司里的中国籍工程师在中国境内设计完成，只要在设计过程中用到了任何“美国技术”（最常见的就是EDA软件，这个是几乎所有设计公司都无法避免的，而且当前三款主流EDA软件都属于法律意义上的“美国技术”），那么其产品就受到美国出口管制约束。也就是说，这家中国公司如果要到台积电去流片，必须获得美国商务部许可。而这个许可，大概率是拿不到的。而拿不到许可，台积电就不会接受其设计文件，因此也就不会为其流片。因此，自家产品性能超过美国设定的红线，对于中国芯片设计公司来说，最直接也是最大的后果，就是失去了台积电的流片渠道。

而所有性能接近A100的国产芯片，无一例外为7纳米设计。没有了台积电还能去哪里？大家自然会想到中芯国际的7纳米。此前华为的Mate 60（传说中）采用中芯国际7纳米的消息对于国产芯片来说自然是一个重大的发展。但中国芯片公司今天面临的不是一个“有没有”的问题，而是一个“什么时间能有”的问题。从消费级的手机芯片到工业级的算力芯片，良率会是一个巨大的挑战（GPU芯片的面积可能上百倍于手机芯片，这意味着同等条件下良率会大大降低）。等到中芯国际工艺跑通，良率稳定，产能提上来，国产芯片终于能够实现接近英伟达A100性能产品本土量产了，英伟达是不是已经发布了传说中的“X100”（传闻中B100之后一代的算力芯片）呢？在大模型一年出两代的世界里，慢，相当于没有。

最后说说赢家。现在手里有足够A800现货（A100更佳）的中国企业，无论是转卖还是自己开发，都是奇货可居，处于极有利的位置。次之，如果现在能够大量囤到H20，也不失为一种可行的过渡方案。从供应端来说，那些早早在中芯国际有7纳米研发优先级甚至是产能预定的国产GPU公司，未来将坐拥一个没有英伟达竞争压力的国内巨大溢价市场，其中暴利难以想象。

因为中国无论如何不可以接受在人工智能发展水平上落后世界前沿两代以上。未来几年很有可能出现各种玩家在“国家战略”支持下，扫尽市场上一切“勉强能用”算力芯片的情况。

算力即国力，谁手上有卡谁发财。

#芯片 #出口管制 #半导体 #中美关系 #中美科技竞争 #评论

本刊载内容版权为端传媒或相关单位所有，未经端传媒编辑部授权，请勿转载或复制，否则即为侵权。

端傳媒的下一程，需要你的守護。今天就成為訂閱會員，支持我們走下去，支持華文世界不可或缺的深度報導和多元聲音。點擊了解更多會員計畫