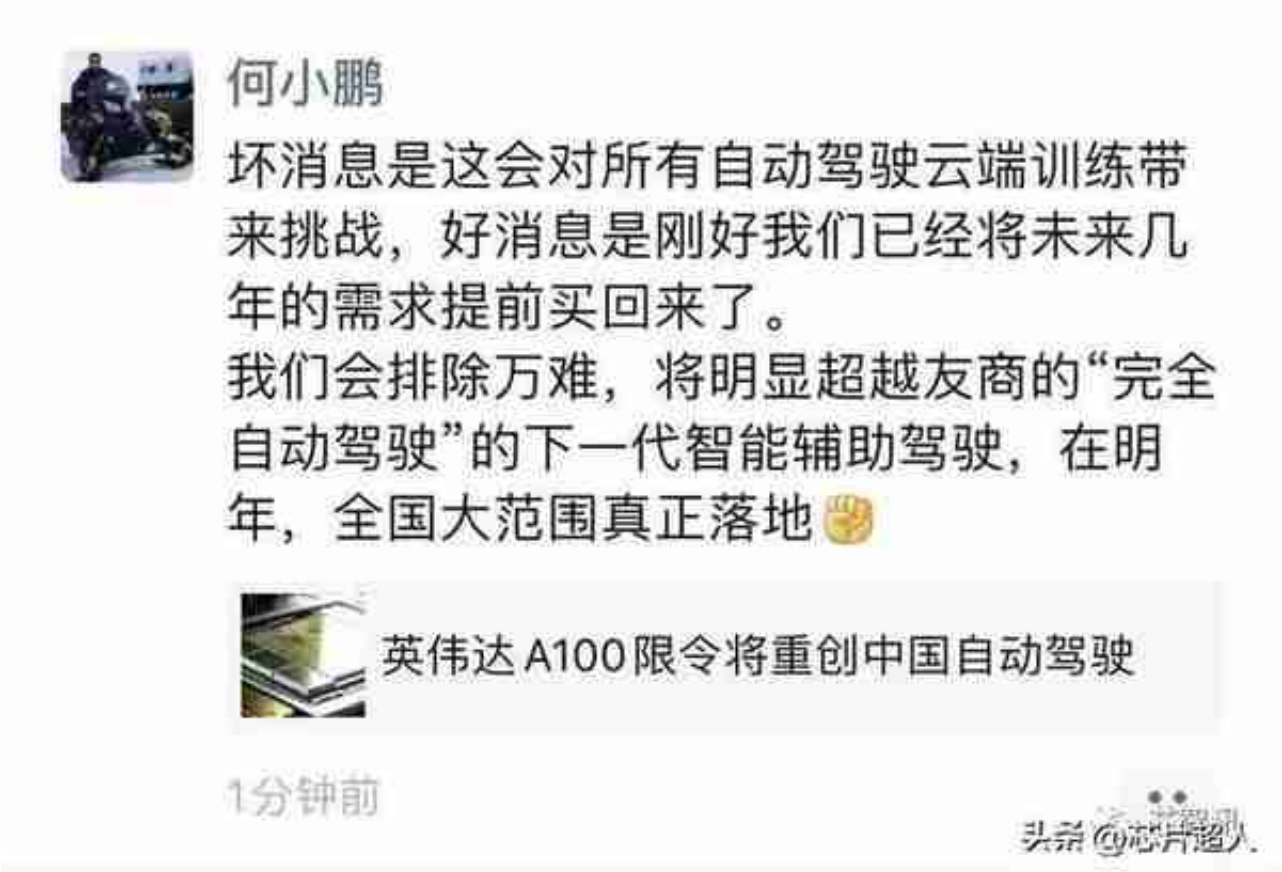


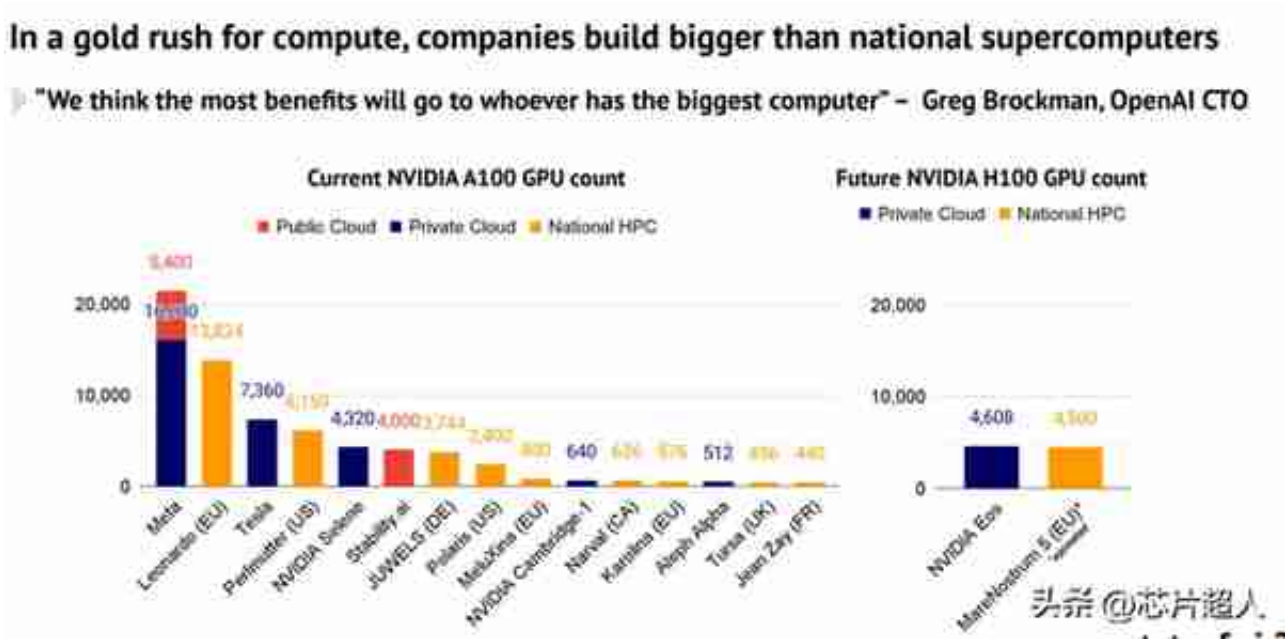
小鹏汽车有提前备货，但禁售令对自动驾驶或其他AI领域的中国企业有多大影响，就不得而知了。在市面上，A100因为变成稀缺品，价格开始水涨船高，从官方的1万美元/枚，约合人民币7万，涨至8万、9万元，快要到10万一枚。即便去年年底英伟达推出A100“阉
割版”（带宽被限制）
——A800显卡，于2022年Q3投入生产，在中国依然遭遇严重缺货。

据了解，A800京东官网定价超过8万元/枚，甚至超过A100官方定价。3月初，有云厂商人士接受财经十一人采访表示，A800实际售价甚至高于10万元/枚，价格还在持续上涨。A800目前在浪潮、新华三等国内服务器厂商手中是稀缺品，一次只能采购数百片。


这代DGX A100 AI系统包含8块A100 GPU

ChatGPT主要就是用英伟达的A100进行训练，这款显卡也成为了最抢手的产品之一。某宝上关于A100
显卡商品的问答中，就有购买者表示自己买来做深度学习，训练模型。

在摩尔定律最好的十年，AI
处理速度提高了100万倍，而到了黄仁勋提出的“黄氏定律”（Huang’s Law）
，从2012年的K20X到2020年的A100，英伟达的GPU推理性能提高到原来的317倍，远超摩尔定律的发展速度。



如上图显示，排名前三位的分别是Met
a（21400）、欧洲超算Leonardo（13824）和特斯拉（7360）。

Meta在去年宣布建造全球最快的AI超算“RSC”，包含16000颗A100
GPU，目的即是为了元宇宙平台
。元宇宙概念是2022年引爆互联网及科技产业的热词，元宇宙本质上是对现实世界的虚拟化及数字化的过程。它本身不是新技术，但却融合了包括5G、云计算、AI、虚拟现实、物联网、人机交互等一大批现有的尖端技术。

来自意大利CINCA研究中心的Leonardo超级计算机使用了近14000颗A100
GPU
，被称为世界上最强大的AI系统。超级计算机多用于国家高科技领域和尖端技术研究，由于其集功能最强、运算速度最快、存储容量最大的优势集一身，在气候、材料学、生物医药、先进制造、航空航天等领域有着广泛的用途，可以模拟大气、气候和海洋，对地质灾害进行预测，也可以对药物研制、生化反应进行模拟，大幅缩短研发时间。

总之，超算是你平常看不见摸不着甚至鲜有耳闻，但却在一个隐秘角落为科技发展发光发热的劳模。

特斯拉在CVPR
2021（国际计算机视觉与模式识别会议）
上公布了内部用于训练
Autopilot与自动驾驶深度神经网络
的超级计算机。这个集群使用了720个节点的8个NVIDIA A100
GPU（共5760个GPU）。

我们知道特斯拉是纯视觉自动驾驶的推崇者，在没有激光雷达提供3D空间数据的情况下，特斯拉仅依靠摄像头提供的2D图像就能完成现有的辅助驾驶系统，这背后是上百万台特斯拉，每天在路上行驶获得的海量图像数据，以及特斯拉为此构建的神经网络模型。

除了这三位外，榜单其余部分均是公有云、私有云和国家超算。

国内企业部分，基于有限的资料，能大量使用A100的大致分为三类：一类是阿里、百度、腾讯（俗称BAT）等云服务商
，另一类是浪潮、联想、新华三等系统集成商
，第三类是像小鹏等自动驾驶车企，但从整体规模来看，海外明显占据上风。

从以上企业所处领域，我们大致归纳出A100等训练芯片主要应用的场景：云计算、超算、深度学习模型训练、自动驾驶、元宇宙、机器视觉等，深入的领域包括：工业、医疗、金融、气候、农业、能源、消费、汽车、半导体等。这些场景和领域往往面临着超大规模的密集型数据、海量存储及高性能计算。

它们都需要不止一
块强悍的芯片，强到连英伟达竟也成为了自己供应商的供应商。

黄仁勋在今年GTC演讲中宣布，新发布的基于 GPU 的计算光刻软件库 cuLitho，用于芯片制造中最复杂、最昂贵的光刻环节，使用它之后计算光刻速度可以提升至原来的 40 倍，光掩膜产能提升 3 至 5
倍，电力消耗减少为当前的九分之一。黄仁勋说，cuLitho 将辅助芯片制程向 2 纳米及更先进迈进。

跟英伟达一起研发该技术的三家公司分别是台积电、ASML 和 Synopsys。台积电帮英伟达代工生产 GPU
芯片，是它最重要的供应商之一。ASML 和 Synopsys分别是全球光刻机和 EDA 龙头，都处于整个半导体产业最上游环节。现在三者都要用英伟达的 GPU 和技术平台。

这意味着能用A100的玩家们并非等闲之辈，能用得到A100的地方也并非是通用化的场景，这就造就了A100这类芯片独特的身份，它的应用范围和使用人群非常的聚拢和突出。

03

真有那么多需求吗？

让我们从海内外终端大厂回到国内现货市场，A100市场价涨至8万、9万，谁会去买单？真有那么多需求吗？

相关从业者对芯世相表示，虽然卖家都在涨价出A100，但成单估计很少。

首先，非刚需，成本高昂，国内能“烧”得动A100的企业寥寥无几。

企业要想玩AI，得经历训练和推理两个环节，一个相当于在校学习，一个是通过所学知识去应对考试，不断精进。没有训练，就不会有推理。

训练
，讲究绝对的算力性能。这里我们分成
两种情况：一种是以GPU为主搭建算力基础，另一种以AI芯片为主。

早些年，进入AI领域的初创企业可以凭借如百度飞浆这类的深度学习开源平台进行研发，但对于生成式AI这类预训练大模型难度极高，大多数创业者哪怕是国际大厂能够将其商业化的也是寥寥无几。

这类模型需要海量的数据、算法、算力的支撑，高投入和高门槛使得很多企业退避三尺。微软直言，给OpenAI打造的超算集群，光建设成本就不止几亿美元。且国内并非所有企业都用得起A100这类的高端显卡，目前国内云厂商主要用的是英伟达中低性能产品如A10，拥有超1万枚GPU的企业不超一只手，拥有1万枚A100的企业只有一家。

除了价格昂贵，维修费也惊人，且保修一般只有一年。对于一般企业而言，如果想要算力的支撑，可以寻求云服务提供商，调用它们的云计算能力，或者去租借高端GPU来完成某个项目的开发。