

1. 历程

顺应时势 创造机遇

英伟达凭借GPU开疆拓土,实现了对图形处理芯片这一空白领域的占据,并保持强势地位至今。

古语说,三十而立。创立于1993年的英伟达,今年刚好30周年,而30岁创立英伟达的黄仁勋则刚度过60岁生日。

黄仁勋在20岁的时候曾对初恋女友也是后来的妻子Lori承诺,“我会在30岁时拥有自己的公司”,1993年2月17日,黄仁勋的30岁生日,真的以4万美元创办了英伟达。

彼时IBM电脑开始了个人电脑革命,个人电脑逐渐从办公室走向家庭,黄仁勋敏锐感觉图片芯片时代即将到来,并将目标锁定在让电脑图像流畅运行的图形芯片,将当时只会运算的个人电脑身为游戏、多媒体的主流消费设备。而在此之前,黄仁勋先后在AMD和LSI工作,期间还在斯坦福大学拿到硕士学位,拥有丰富的产品研发、市场销售的经验以及相关的行业资源。

黄仁勋在LSI公司就职时,在设计部门待了两年就要求调到销售部门。“从工程部转到销售部,这是我曾经做出的最佳职业选择,我逐渐学会了产品的设计开发如何与市场结合。因为我意识到,消费者并不关心你从哪家商学院毕业,他们只关心一件事:你的产品对他有什么好处。”黄仁勋表示。

而在AMD工作的两年时间,他专注研发工作,这在广大的网友眼中则被戏称为“金牌卧底”。不过,创业过程却没有那么多浪漫,英伟达和黄仁勋很快迎来了创业生涯的第一个挫折。

当时图形芯片赛道刚刚兴起,已经有二三十家同类公司开展竞争,当时很多风投和咨询公司的分析师都认为图形芯片市场几近饱和,且无序竞争、前景一般,但黄仁勋还是义无反顾地走上了这条路。不过英伟达的前两款产品NV1、NV2,因偏离主流技术且成本高、无性能优势均告失败,公司濒临破产。

好在游戏公司世嘉及时带来700万美元的游戏主机芯片合同,让英伟达逃过一劫。不过在经历了一年多的开发后,英伟达和黄仁勋意识到设计架构错了,继续开发的结果是产品会和市场主流的Windows不相容。

当时黄仁勋进退两难,完不成合约,公司破产;完成合约,产品融不进主流市场,迟早也要关门。于是黄仁勋找到世嘉的执行长,向他解释了开发路线的错误。世嘉为英伟达的错误买了单并延长了3个月的运营周期,就在这3个月时间里,英伟达终于开发出符合微软显卡标

准的Riva128,打开市场的同时,知名度也由此打响,从而拯救了公司命运。

“坦承错误、立即改正并寻求帮助”成为黄仁勋在这次创业经历中的最大感悟,而3个月的顽强奋斗起死回生,也见证了英伟达的求生意志和顽强坚持。

此外,英伟达可能更应该感谢的是,赶上了对图形图像处理产生巨大需求的互联网时代。英伟达凭借GPU开疆拓土,实现了对图形处理芯片这一空白领域的占据,并保持强势地位至今。

这些只是英伟达踩中风口开始。很快黄仁勋意识到,自家生产的GPU应该有更广阔的应用市场。于是在2007年,英伟达推出CUDA加速计算技术,可利用图形处理器进行通用计算处理工作,英伟达开始着力构建GPU的软硬件生态。

现在回头看,这是英伟达建立AI时代领先优势的关键之举,不过对于当时的英伟达来说,这个超前的想法却带来了巨大压力。因为当时CUDA开发和运营成本很高,却没有明确的商业盈利方向,而且产品开发成本增加了,但产品售价却没变。英伟达的利润急剧下滑,股价陷入低迷,市值仅仅维持在10亿美元上下。这也遭到了股东和外界的众多质疑,希望英伟达能更专注于提升盈利能力。

不过,英伟达扛住了企业盈利和外界环境的各种压力,坚持到了收获的时节。英伟达的这个软硬件生态优势有多大?“院士与高中生的差别。”芯片工程师陈锋告诉记者,“CUDA跟Windows和安卓系统很像,已经足够成熟和方便,生态太强了。其他AI芯片公司也可开发类似CUDA的工具,但根本没人陪你玩。”

GPU并行计算的威力也开始显现。2010年,谷歌AI项目Google Brain的AI大佬吴恩达面临课题障碍:训练神经网络AI的算力不够了,最终其并联了48块英伟达GPU就获得用其他1.6万块CPU才有的理想效果,解决了这个难题。2012年斯坦福大学mageNet竞赛上,2块英伟达GTX580 GPU训练出来的AlexNet计算机视觉模型,以压倒性优势击败谷歌、微软等一众科技巨头,拿下挑战赛冠军,更是让英伟达一战成名。

这让无数人意识到,GPU对于人工智能开发的重要性和广泛应用空间。黄仁勋更是将2012年视为

而这么做的结果就是前面说的,英伟达放弃了过去十多年最大的蛋糕,却获得了今天AI时代的爆发。这让人感慨,有舍必有得。

“‘撤退’对聪明且成功的人来说并不容易。然而,战略性的撤退、牺牲,决定放弃什么是成功的核心。”黄仁勋感慨,“我们的放弃获得了回报,我们创造了一个新的市场——机器人技术。拥有神经网络处理器和运行AI算法的安全架构。”

但放弃不是舍弃更不是躺平,只是为了更好地聚焦。芯谋研究总监李国强对记者表示,英伟达的成就源于它的专注,其长期专注GPU领域,又有AI生态的优势,这两个方面最终让英伟达成为全球芯片产业王者。

其实各厂商的AI芯片的性能差距并不很大,也可以追赶,拉开差距的关键在于英伟达的CUDA计算架构及软硬件生态,CUDA让只做3D渲染的GPU得以实现通用计算功能,应用领域得以从游戏(图形渲染)向外扩展至高性能计算、自动驾驶等多个领域。

作为一个免费、强大的并行计算平台和编程模型,CUDA的低成本和兼容性成为其最重要的卖点之

英伟达万亿帝国背后

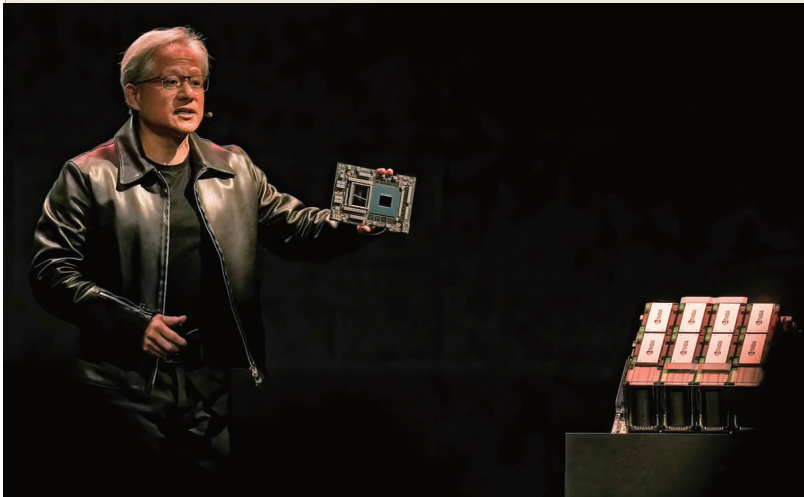
编者按/ ChatGPT的爆火和AI(人工智能)时代的到来,让芯片巨头英伟达(NVIDIA)和它的掌舵人黄仁勋成为最大赢家,原因是英伟达的高性能AI芯片及软件是AI时代的燃料和硬通货。对此,英伟达创始人黄仁勋霸气宣称:“英伟达不仅处在AI浪潮里,而且是浪潮得以成型和推进的原因,我们是全球AI的发动机。”

AI的产业东风让英伟达身价水涨船高,近日英伟达股价一度飙升至419.38美元/股,市值突破万亿美元。英伟达成为全球首家市值突破万亿美元的芯片公司(是芯片巨头英特尔市值的八倍)、全球首家由华人创立的万亿美元市值公司,并成为当前市值仅次于苹果、微软、Alphabet、亚马逊的美股五大科技股。

英伟达是一家以设计显卡和芯片为主的半导体公司,创立于1993年,美籍华人黄仁勋是创始人兼CEO。当然,对于国内很多游戏玩家来说,他们更熟悉的是N卡(电脑显卡)和老黄(黄仁勋的昵称)。不过随着人工智能时代的到来,英伟达官网的slogan早已变成了“人工智能计算领域的领导者”。

作为AI芯片的王者,英伟达市场份额高达91.4%,而排名第二的对手AMD,市场份额仅为8.5%。在深度学习GPU(显示芯片)市场,英伟达更是占有95%的市场,显现出英伟达在这个领域的统治力。美国银行最近一份报告称,英伟达是“AI淘金时代的卖铲王”,并引用了19世纪50年代淘金热期间出售铲子的商人比真正的淘金者更赚钱的观点。

从圣何塞一家不起眼的快餐厅诞生,到万亿美元市值的芯片巨头,英伟达崛起背后有哪些方面值得借鉴和参考? 本期商业案例,《中国经营报》记者重点研究万亿英伟达草根逆袭的背后故事。



2023年5月29日,台北,英伟达创始人CEO黄仁勋在COMPUTEX大会上公布了一批与人工智能相关的新产品和服务。

英伟达的重要转折点,称这一年是“人工智能发展标志性的一年”。此后黄仁勋便在各种场合为AI站台,成为人工智能坚定的鼓吹者。

说起来,英伟达还是ChatGPT热的重要推手之一。2016年,英伟达推出了全球首款一体化深度学习超级计算机DGX-1,当年8月,黄仁勋非常前瞻性和慷慨地把首台DGX-1直接捐给了ChatGPT的开发公司OpenAI。要知道,OpenAI当时只是一家成立不到一年、名不见经传的非营利性人工智能公司,而当时英伟达手里正积压着100多家公司的DGX-1订单。

如今,英伟达的这项超前“投资”正在收获超额回报。AI算力需求的迅速升温,让英伟达的数据中心业务成为公司增长的第一大动力。过去6年,该公司的数据中心芯片业务急速增长了30倍。

一、同时CUDA有着丰富的社区资源和代码库,能为编程提供良好的支持。能作为GPU的核心供应商,基于CUDA架构形成的高度黏性的开发者软件生态,英伟达完成了由传统GPU供应商到平台化公司的蝶变。

一位国内GPU厂商工程师告诉记者,经过多年的积淀,英伟达的CUDA已有近400万开发者,形成了很高的生态壁垒,而软件生态正是英伟达相对于AMD、英特尔等的最大优势。

而且英伟达经历了移动端芯片开拓失利并退出后,并没有停止Tegra处理器的研发,而是改变产品定位,将Tegra处理器运用在智能汽车、智慧城市和云端服务上,并由此初步奠定了英伟达“两产品条线,四市场”的商业模式。两产品条线包括了英伟达传统产品GPU和Tegra处理器,而四市场则包括了游戏、企业级、移动端、云端。

这种兼顾激情和理性、原则性和灵活性的做法,最大化减少了退出、转向的成本和风险。

在李国强看来,英特尔、AMD都是横跨CPU、GPU和FPGA的龙头,而只有英伟达长期专注于GPU。

近些年火热的元宇宙和虚拟货币挖矿潮,英伟达都是最核心的玩家和受益者之一。因为“挖矿”基础动力就是英伟达的GPU,“挖矿”的显卡一度占到占英伟达2021年第一财季总体销量的25%。

甚至当下火热的新能源汽车和自动驾驶浪潮,英伟达也抢下了一大块蛋糕。英伟达2024财年第一财季财报(第一财季截至2023年4月30日)显示,英伟达负责自动驾驶芯片和软件的汽车部门,上季度营收暴增114%至2.96亿美元,是英伟达当前增长最快的业务。在自动驾驶上,当前国内大部分高端车型都选择了英伟达的Orin芯片和多类型传感器的组合来实现辅助驾驶功能。

英伟达几乎踩中了每一次时代浪潮,并最终借此登上了万亿市值的巅峰。

只有足够聚焦才能足够专业。图像处理器市场产品逐步细化,主要分布在游戏玩家、企业级、平板电脑和移动端用户,不同客户的需求差异化明显,针对不同市场,深度聚焦的英伟达推出了对应的产品方案。比如英伟达面向企业市场的产品就包括用于工作站的Quadro,用于高性能计算服务器的Tesla和用于企业VDI应用程序的GRID等。

游戏玩家们常常亲切地喊黄仁勋为“皮衣大佬”和“老黄”。因为英伟达能精准灵活地推出不同档次的显卡,且对不同显卡性能精准阉割,每一阶显卡,在价格和性能上只比次一阶高一些,这种错落有致的各阶显卡布局,也被叫做“老黄刀法”,被网友誉为“黄氏刀法,不差毫厘”。只有足够专业才能赢得足够信赖。英伟达的GPU得以成为包括游戏玩家在内众多行业客户的优选,某种程度上,英伟达已成为GPU的代名词。

正是长期专注和高生态壁垒,奠定了英伟达芯片帝国的霸业。在GPU领域,曾经的王者3DX、ATI陆续被收购,唯有专注的英伟达屹立不倒。

3. 坚持

只要活着 不断奔跑

作为GPU领域的开创者,黄仁勋也提出了知名的“黄氏定律”:显示芯片每6个月性能提升1倍,比摩尔定律提速3倍。

不过当大家都对英伟达GPU+CUDA的强大生态钦羨不已的时候,英伟达已在为下一代谋篇布局,并拿起了对自己革命的刀。

2020年上半年,英伟达以69亿美元对价收购以色列网络芯片公司Mellanox Technologies,并于同年推出BlueField-2 DPU,将其定义为继CPU、GPU之后的“第三颗主力芯片”。

与非网资深行业分析师张慧娟对记者表示:“如今数据量越来越大,传统的CPU处理已经跟不上数据爆发的节奏,尤其是短视频、视觉类应用,数据量呈现指数级增长,DPU应运而生。

英伟达也在构建全新的软件生态,其于2021年推出DOCA。英伟达简单直接地解释道:“DOCA之于DPU,正如CUDA之于GPU。”

“如果说CUDA是GPU的灵魂,那么DOCA就是DPU的灵魂。就如同手机,假如没有丰富的App,就只能打电话、发短信。正是因为有了软件生态,让开发者开发出各种各样的应用,手机功能才如此丰富。”张慧娟表示。

即使一路遥遥领先时还一直在奔跑的英伟达和黄仁勋,谁又能赶超呢?实际上,英伟达和黄仁勋就是在一路狂奔中赶超对手的。

1963年出生于中国台湾的黄仁勋,9岁便被父母送到美国的叔叔家生活。彼时美国种族歧视严重,而黄仁勋在一所问题少年聚集的乡村学校读书,因为年纪小,个子小被安排每天扫厕所,在这种复杂的生存环境成长和与背景各异人群相处中,黄仁勋从小就培养了积极适应环境的能力和斗志昂扬的战斗意志。

正如半导体芯片领域有摩尔定律,作为GPU领域的开创者,黄仁勋也提出了知名的“黄氏定律”:显示芯片每6个月性能提升1倍,比摩尔定律

提速3倍。

“黄氏定律”的核心就是迭代提速和全力升级,这更像是一个自我实现的预言。创业伊始,英伟达便开启了“高速奔跑”模式,其以“三团队-两季度”的研发运营模式,每6个月推出一次新产品,领先市场1~2个研发周期。按这种套路出牌的结果是,累趴下了自己,但也把同行卷出了局。曾经强大无比的竞争对手要么倒下,要么被收归麾下。

1996年,英伟达推出NV3系列的Riva128芯片,在性能方面甚至优于英特尔于下一年推出的i740;1999年,英特尔宣布推出独立显卡芯片组业务;1999年,英伟达推出全球第一颗GPU GeForce256,全面超越当时行业先驱3dfx,最终3dfx宣布破产并被英伟达收购。

在技术和性能领域的提升创新方面,英伟达还通过不断收购、技术开发以及广纳人才,进一步开拓市场,增强自身实力,自驱奔跑式地保持着市场领先。

就在英伟达登上万亿市值巅峰的同时,黄仁勋正忙着在台北国际电脑展上发表主题演讲,并宣布推出新一代“杀手铜”超级计算机DGX GH200,这被认为是当今全球性能最强的运算器产品。而就在两个月前,英伟达刚发布了AI云服务DGX Cloud,通过云厂商平台,向用户提供训练生成式AI大模型及其他AI应用的算力月租服务。

显然,在外界看来,如今的英伟达已足够成功,但英伟达的野心并不止于此,它希望借生成式AI的东风,通过云部署的方式触及更多中小企业,不断扩大AI算力的服务市场。

黄仁勋的大半生,一直没有停歇。60岁的他,称自己暂时还没有退休的计划,要再领导英伟达30到40年,直到90岁左右,届时会以机器人的形式继续工作。

观察

坚持在AI金矿边上卖“铲子”

1956年,人工智能的概念首次提出,迄今已逾六十载。六十多年来,AI历经了从爆发到寒冬、再到野蛮生长的历程,AI热潮也几经沉浮。2022年,ChatGPT的火爆出圈,让AI行业再度迎来新的拐点。

在众多业内人士看来,这一轮生成式AI的浪潮,英伟达是最赚钱的,AI大模型时代,英伟达将再次进入狂飙时期。作为在AI金矿边上卖“铲子”的商人,英伟达已经牢牢扼住了AI时代的算力命脉。

回顾英伟达三十年崛起的历程,其不断将筹码押注在人工智能的举动一度被认为过于疯狂,但英伟达的每一次决策都堪称果断,每一项举措又足够灵活,接连放弃手机芯片市场和矿机市场,最终在人工智能领域收获成功。而积累多年的CUDA生态也成了英伟达最深的一条护城河,支撑起了万亿美元的市值。

在黄仁勋的领导下,英伟达有了鹰的敏锐和狼的嗅觉,有了猎豹的速度和狮子的野性,它可以果决舍弃,也可以柔软求生,更有着一股非同一般的求生、求胜意志。这样的英伟达几乎是无敌的,它的竞争对手们都已感受到了。

但是,当一个商业帝国如日中天的时候,或许也到了最危险的时刻。从柯达到诺基亚,再到

倒在自己面前那一个曾经强大无比对手,看过和经历过它们的兴衰成败,黄仁勋应该心有戚戚焉。

在英伟达站上万亿美元市值巅峰的同时,有多少惊叹倾羨,就有多少批评质疑,英伟达的业绩是否能够支撑起万亿市值并持续保持?

对英伟达而言,生成式AI热潮带来的商机到底有多大和多持久,还有待观察;从市场竞争的角度来看,黄仁勋的对手也在加速追赶的过程中,都想来争夺AI芯片爆发的美味蛋糕。

据报道,云计算前五大厂商亚马逊AWS、微软Azure、Google Cloud、阿里云和IBM都在做自己的AI芯片。而且中国内地AI芯片厂商的崛起,也会对英伟达形成冲击。

其实,在ChatGPT火爆前的2022年,由于挖矿退潮和消费电子的低速等因素,除了数据中心业务,英伟达其他业务表现欠佳,股价也持续下跌。实际上,英伟达在其30年的创业历程中,已经历了多次沉浮乃至濒临破产,但只要没有倒下,黄仁勋和英伟达总能在下一个风口高高飘起。

外界的喧闹、赞美、质疑属于外界,在这场无尽的长途奔跑中,30岁的英伟达和60岁的黄仁勋选择了继续奔跑。

本版文章均由本报记者吴清采写