科技周期探索之七

2016-2030年：通用人工智能时代的到来

中性

核心观点

路线的转变：从CPU到GPU的切换

2016年全球的移动互联网渗透率已经超过了50%，这代表的其高速增长的时间已经过去；此时英特尔公司放弃了引以为傲的“Tick-tock”战略，CPU在终于在算力提升的路上严重受阻。这这样的背景下，科技界都在寻找新方向，即能够接过CPU接力棒的技术。凭借早年显卡的积累，以及CUDA架构的提出，英伟达GPU逐渐成为通用计算芯片，它替代CPU成为引领算力进步的新宠儿。

尽管英伟达在人工智能芯片上的单芯片算力从2012年开始用了10年的时间翻了1000倍，但是以提升功率与价格的方式实现的，我们测算最近10年全球每GFLOPS的复合成本降幅大约在25-35%之间，这一降幅略低于摩尔定律的要求。

三大算力应用：比特币、云计算、新能源车

比特币的出现大大拉动了全网算力的提升，在14年的时间里，全网算力增加了3万亿倍（3.2*10^12），相当于每年复合增速6.8倍。目前比特币挖矿的耗电量相当于全球排名第20名左右的国家用电量。

单CPU算力提升受阻,云计算成了继GPU之外的又一解决方案。到了2016年，几乎所有的大公司在云计算的部署都已完成。2024年，云计算市场规模达到了6760亿美元，2016-2024年复合增速达到了25.2%。

从算力角度，新能源汽车的第一特征可能不再是汽车，而是一台“大号的、行走的计算机”，从Model-S开始，新能源车的发展进入到了快车道。中国在新能源车的普及上遥遥领先全球，2023年新能源车占总销量比重高达38%，欧洲为21%，美国仅为9.5%。单车载芯片算力目前达到500-1000T，预计2030年将达到5000T。

大模型的出现：AI翻开了崭新的一页

2017年，有关Transformer架构的论文发布，随后谷歌的BERT模型与OPENAI的GPT模型发布，但初期并未受到广泛关注。2022年GPT3.5成为分界点，它让科技界看到了“大力出奇迹”的千亿参数级别的大模型效果可以如此强大。随后万亿参数级别的GPT 4、以及多模态SORA模型的出现，为大模型的发展打开了更广阔的空间。

如果将模型参数与人类的神经突触对标，那么大约到100万亿参数的模型可以实现AGI（通用人工智能），马斯克、黄仁勋、OPENAI前员工大约预测了这个时间在2027-2029年。

鉴于AI Agent的开发门槛越来越低，LLM能力越来越强，它或将成为下一个风口。就如同移动互联网时代的APP，互联网时代的网站，计算机时代的应用软件，AI Agent或将走出下一批大公司。

风险提示：地缘政治的不确定性，美联储降息幅度的不确定性，部分行业竞争格局的不确定性。

行业研究·海外市场专题

美股

中性·维持

证券分析师：王学恒
010-88005382
wangxueh@guosen.com.cn
S0980514030002

市场走势

资料来源：Wind、国信证券经济研究所整理

相关研究报告

《美股市场速览-科技巨头带领大盘上涨》——2024-12-08
《美元债双周报（24年第49周）-美债利率高位回落，12月降息预期升温》——2024-12-03
《美股市场速览-大盘延续涨势，消费板块领先》——2024-12-01
《美股市场速览-大盘反弹，传统行业风格延续》——2024-11-24
《美元债2025年展望:非典型降息周期下的美元债投资策略》——2024-11-24

内容目录

图表目录

在本篇报告中，我们梳理 2016 年以来的后移动互联网时代的发展，以承接上两篇报告《2002-2016 年：移动互联网的大时代》、《案例篇：移动互联网时代的十倍股与百倍股》。我们将 2016 年至 2030 年称之为“人工智能时代”。

但这里会有疑问：“人工智能”这个概念，如果用于区别“移动互联网”，读者都觉得不难理解。但对于未来，我们能说人工智能在 2030 年之后就不发展了吗？如果到那个时候 AI 还会进步与发展，应叫什么时代呢？实际上从某种意义上说，用“人工智能”这个词去展望未来几十年似乎都不为过，那么仅用它来称谓短短的 14-15 年的一个科技周期，合不合适呢？

所以，报告写到当代，我们也遇到了一个回避不了的难题：回顾过去，慷慨激昂；展望未来，如履薄冰。

从 2016 年以来，我们都经历了什么？2016 年谷歌的 Alpha Go 问世，这是 AI 的一个里程碑式的事件。2016 年前后，正是亚马逊云计算扭亏为盈后的一年，随后几年迎来了高速的发展。同时，微软、谷歌、阿里巴巴、腾讯等巨头们也都大力发展云计算，云计算似乎是过去一些年的主线，但又不尽然；

在并行运算领域，英伟达不停地更新架构，从 2016 年的 Pascal、到 2017 年的 Volta、到 2018 年的 Turing、到 2019 年的 Ampere、到 2022 年的 Hopper……这样来看，除了英伟达，还有 AMD、谷歌、ARM 以及中国诸多科技企业也都在加码 GPU，这也是一条主线；

2019 年开始，特斯拉、中国的造车新势力、比亚迪等诸多企业，开始推出物美价优的新能源汽车，尤其在中国，新能源汽车在汽车销量占比迅速提高，目前已经突破了 40%。随着新能源汽车的普及，汽车智能化、无人驾驶也变得越来越普及。2023 年华为无人驾驶发布，2024 年百度在武汉大规模试商用 Robotaxi，这样看，新能源汽车、无人驾驶也是一条主线；

2022 年底，ChatGPT 的诞生，大语言模型横空出世。人类第一次不同于历史，见证了如此流畅的语意表达。图灵测试瞬间从“可能遥远的将来”变成了“就是现在”。这让人类欢呼雀跃，以大模型为代表的 AI 技术革命又进入到了崭新的时代，除了聊天，大模型绽放出了更强大的能力：图像处理，文生视频，行业应用……让我们不知道它的边际在哪里。这样看，主线又回到了 AI。

因此，我们决定依然将现在所处的时代称之为“人工智能”时代。尽管，未来或许 AI 还会有新的变化，但或许到那时再来修正今天的这个称谓也是可行的——因为无论怎么看，唯有人工智能才能最广泛地、最恰当地匹配当代的所有主线。

路线的转变：从 CPU 到 GPU 的切换

2016 年后移动互联网增速开始放缓

在一个新的浪潮里，产量曲线（或者销量）会有如下特征：

1、它总体呈现的是一种高斯分布；

2、在早期，随着时间的推移，产量开始陡峭增加，行业进入加速期；随后的几年里，该行业呈现非常景气的局面，收入、业绩、估值均在快速提升；

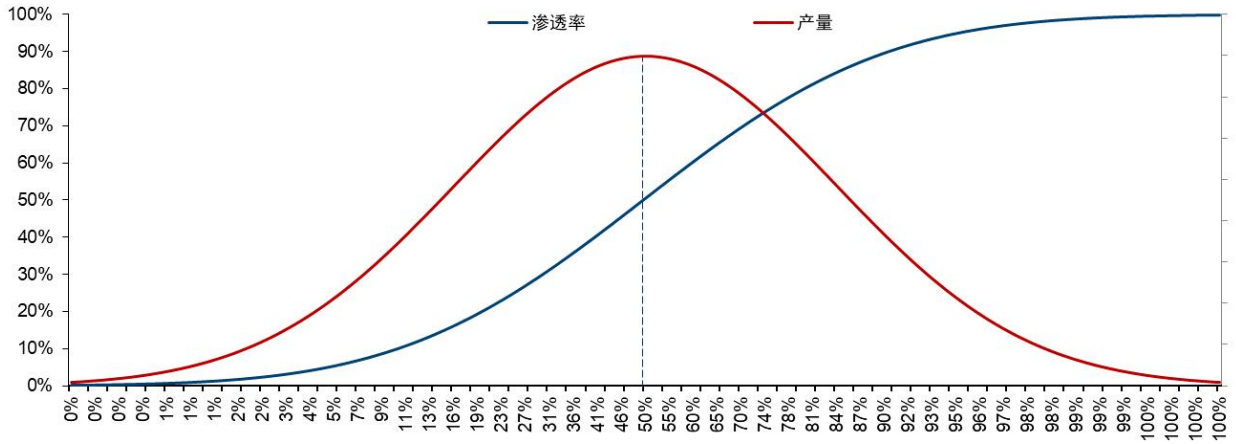
3、走过了最快增速的几年，行业增速开始放缓直到增速为零；此时行业渗透率达到了 50%，即一般的目标价人群已经使用了该产品/服务；此时行业的竞争变得激烈，行业中绝大部分公司都面临着增长乏力的问题，估值早在一两年以前就开始回落了；

4、行业渗透率超过 50%以后，由于产量下滑，诸多平庸的企业日子不好过，收缩业务，几家龙头公司的份额逐渐扩大，它们成了“剩者为王”（即是剩者，也是胜者）。尽管渗透率还在稳步提升，但是估值已经没法给到几年前那种上百倍、几十倍的水平了。

我们来思考一个问题。如果站在划分周期的角度，我们能否在渗透率曲线上寻找一个位置（虽不一定特别严格），来作为时代与时代之间的分水岭呢？

例如 1844 年电报就发明了，在随后的一百多年里，它一直存在。在中国，1871 年上海开始了电报业务，一直到 2021 年上海最后电报系统退网。150 年的时间里，电报业务成了“永不消逝的电波”，那我们能说，过去的 150 年里是“电报时代”么？问题在哪里？你会说：明明过去的 150 年里出现了那么多技术，为什么只说电报时代呢？还有电话，BP 机，大哥大，手机，电脑啊...

图1：产量与渗透率



资料来源：Wind，国信证券经济研究所整理

是的，没有一个时代仅包含一种产品或技术，它一定是多个产品、多种技术的结合。技术周期是我们抽象出来的，是试图发现推动人类科技进步那个最大的浪潮或者最主要的科技驱动力，而大浪潮下一定有诸多小波浪，且有的在先，有的在后，我们期望要抓住主要矛盾以认清这些规律。回头看，每个时代的大浪潮只有一个，但不容易的是，身处浪潮之中的我们，经常会被小波浪所困（尤其是在浪

潮的早期，哪种技术会成为大一统并不是显而易见的，就像我们提到的过去几年的科技主线似乎有很多条一样）。

回顾互联网与移动互联网，我们发现，50%，大约可以是一个浪潮阶段性终结的标志。例如，美国互联网渗透率到 2002 年首次超过了 50%（达到了 59%），而 2002 年也是我们划分的互联网时代的终点。同样，全球移动互联网的渗透率并不统一，但我们倾向于用“移动宽带”（ITU 口径）这个“短木板”来替代移动互联网渗透率，它于 2016 年首次超过了 50%（达到了 52%），因此我们可以也将 2016 年作为移动互联网时代的终点。需要说明的是：

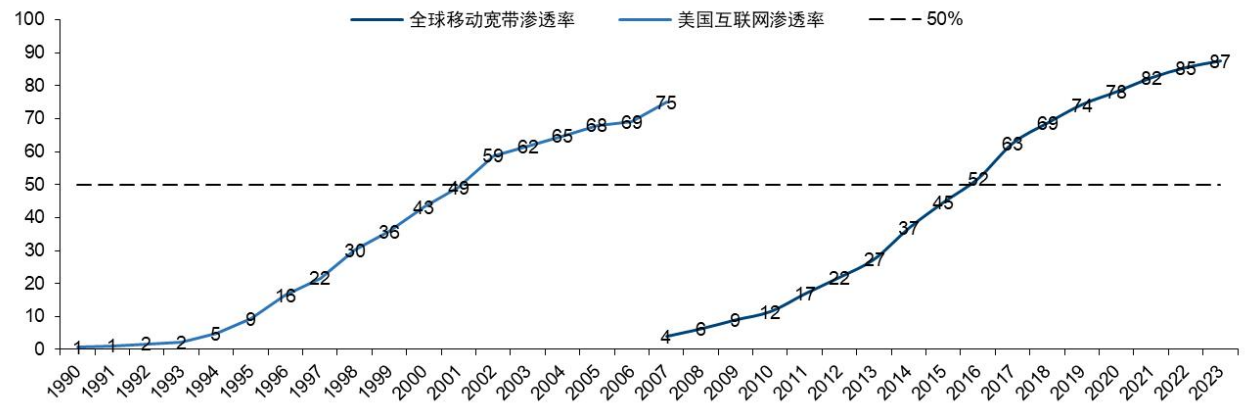
1、这种终点：只是我们在划分科技周期中的思考，因为我们划分周期的目的是指导投资。在《经济周期随笔》系列，“如何划分基钦周期”中，我们也提到过这种思想，我们划分周期的目的是什么？指导投资？指导企业？还是指导宏观调控？不同的立场与出发点，会使得我们划分周期的结果是有差异的；比如，当你说，互联网时代哪里结束了啊？我们至今还在用啊。那请你回顾一下思科、美国在线、雅虎、世通公司.....这些科网泡沫的例子，看看这些公司是否还在创新高？或者看看它们有多少家公司从高点跌去多少幅度，甚至大面积退市。所以我们在划分科技周期的考虑是期望解释与指导投资，这是前提，不同的讨论前提必然有不同的结果；

2、50%这个分水岭，是实业中“成长到成熟”的标志，也是资本市场中“盛极而衰”的标志。大量的案例表明，成熟期的企业并不能用成长期的高估值去衡量。但这并不意味着行业的衰败或者消失，就像电报的例子，它还可以存在很多年；

3、一个旧时代的终点，也必然伴随一个新时代的起点。例如，2002 年，互联网大时代的结束，但也意味着移动互联网时代的开始；2016 年，移动互联网时代的结束，但也意味着人工智能时代的开始；

4、并非每个时代都有一条清晰的渗透率曲线。现实不见得是完美的。例如，有的时候渗透率曲线可能会时间更长（PC 时代），有的时候可能会突然加速（互联网时代）。以及，过程中会存在不同产品的渗透率曲线混淆视听。比如本时代你一定能够找到新能源车的渗透率曲线，但这个时代是新能源车时代吗？用新能源车来概括人工智能，明显是有些以偏概全了。这是个问题，我们先放在一边，等到我们在后续的报告中集中讨论。

图2：互联网/移动互联网两段渗透率曲线对比



资料来源：世界银行、ITU，国信证券经济研究所整理

回到 2016 年，移动互联网时代告一段落，新时代大幕徐徐拉开。

2016 年英特尔放弃“Tick-Tock”

英特尔在移动互联网时代中未能像苹果、谷歌那样成为主角，但它依然在推动 CPU 的技术进步。

早在 2000 年，高端 CPU 需求增长放缓，AMD 凭借新产品在低端和中端处理器蚕食了英特尔的市场份额，英特尔在其核心市场的主导地位大大削弱。2004 年和 2005 年，AMD 又对英特尔提起了与不正当竞争有关的诉讼索赔，这段时间是英特尔的低谷期。

2006 年，英特尔发布酷睿架构（Core），该产品系列被认为是处理器性能的一次飞跃，一举夺回了英特尔在 CPU 领域的领导地位。为了保持领导者地位，2007 年，英特尔提出了 Tick-tock 模式（Tick-tock model）。所谓“Tick-tock”，就是锁定摩尔定律，期望每 2 年就将 CPU 的制程升级一次：像钟摆一样，“Tick”1 年，提升 CPU 的制程工艺；“Tock”1 年，不提升制程工艺的前提下，通过改善设计来优化性能。所以，“Tick-tock”相当于是是制程工艺“tick”1 年，设计优化“tock”1 年，“Tick-tock”就是一个完整的周期，合计 2 年。

“Tick-tock”在初期运行的很好，如 2007 年 11 月 45nm 制程工艺下“Tick”1 年，紧接着 2008 年 11 月“Tock”1 年；2010 年 1 月 32nm 制程工艺下“Tick”1 年，紧接着 2011 年 1 月“Tock”1 年；2012 年 4 月 22nm 制程工艺下“Tick”1 年，紧接着 2013 年 6 月“Tock”1 年；2014 年 9 月 14nm 制程工艺下“Tick”1 年，紧接着 2015 年 8 月“Tock”1 年。

按照这个节奏，到了 2016 年，应该是在 10nm 制程工艺下的“Tick”，但不幸的是，2016 年英特尔卡壳了。而且这一卡壳，不是 1 年，不是 2 年，而是漫长的 7 年。因此，2016 年 3 月，当英特尔意识到自己无法交出答卷之时，它宣布放弃“Tick-tock”生产周期，转而采用一种可在更长时间里使用相同尺寸晶体管的工艺。

表1：英特尔的“Tick-tock”模式

模式	制程工艺	架构名称	时间
TICK	45nm	Penryn	2007/11/11
TOCK	45nm	Nehalem	2008/11/17
TICK	32nm	Westmere	2010/1/4
TOCK	32nm	Sandy Bridge	2011/1/9
TICK	22nm	Ivy Bridge	2012/4/29
TOCK	22nm	Haswell	2013/06/02
TICK	14nm	Broadwell	2014/9/5
TOCK	14nm	Skylake	2015/8/5
	14nm	Kaby Lake	2017/1/3
	14nm	Coffee Lake	2017/10/5
	14nm	Whiskey Lake	2018/8/28
	14nm	Skylake	2018/10/8
Optimization	14nm	Coffee Lake	2018/10/8
	14nm	Cascade Lake	2019/4/2
	14nm	Comet Lake	2019/8/21
	14nm	Cascade Lake	2020/2/24
	14nm	Cooper Lake	2020/6/18

资料来源：维基百科，国信证券经济研究所整理

“Tick-Tock”的跳票，也是摩尔定律悄然变化的一个缩影。

1965 年，摩尔提出了摩尔定律，当时的预测是单位面积晶体管数量将以“每年大

约两倍的速度”增长（后来又被摩尔的同事修正到 18 个月）。但到了 21 世纪，有一些端倪显示，这个周期开始变得不稳定，摩尔定律开始遭受挑战：

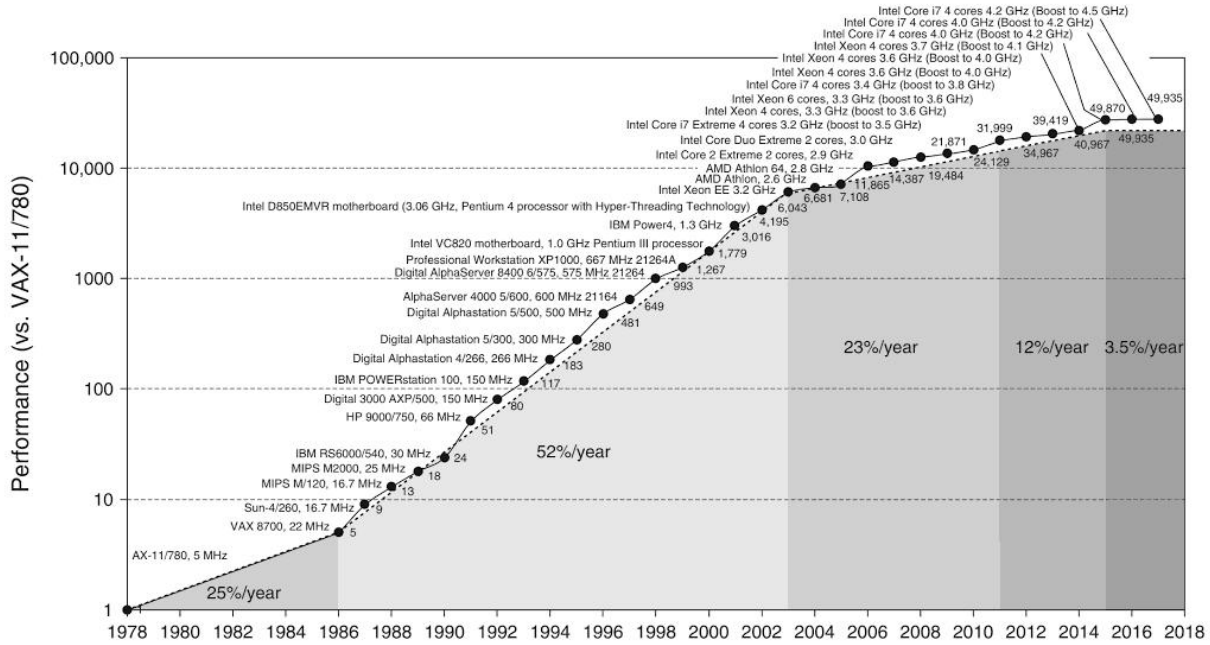
挑战一是硅物理极限正在被触及，随着晶体管变得越来越小，量子效应和制造挑战变得更加明显。例如登纳德缩放定律（Dennard Scaling）——该比例假定功耗将与晶体管的面积成比例减少（电压和电流都与长度成比例）。根据登纳德缩放比例，晶体管尺寸每代技术将缩小 30%（0.7 倍），从而将其面积减少 50%。这将减少 30%的延迟，从而将工作频率提高约 40%（1.4 倍）。为了保持电场恒定，电压将降低 30%，能量减少 65%，功率减少 50%。因此，在每一代技术中，晶体管密度都会翻一番，芯片速度提高 40%，而功耗保持不变。但由于在小尺寸下，电流泄漏会带来更大的挑战，也会导致芯片发热，从而造成热失控的威胁。

挑战二是制造成本的增加。随着工艺节点越来越小，每一步微缩都要求更昂贵的制造设备，如极紫外光刻（EUV）技术，这使得新晶圆厂的建设成本飙升。良品率也是巨大的挑战，英特尔在 10nm 制程上良品率就花了很长时间来改善。

挑战三是功耗增加。除了登纳德缩放定律提到的电流泄露问题，晶体管数量的增加必然导致功耗增加，而电池技术的发展这些年来始终慢于摩尔定律的增加（例如智能手机待机时间大约只有 1 天，笔记本只有几小时），这反过来影响芯片的稳定性和寿命，以及设备的整体体验。

挑战四是设计的复杂度。更小的制程意味着芯片设计的复杂度呈指数级增长，需要更复杂的电路设计和验证工具，以及更多的研发时间和成本。

图3：处理器性能提升速度放缓



资料来源：eetchina，国信证券经济研究所整理

以下是一些典型事件：

1、单核 CPU 的性能提升速度已显著放缓：单核性能在 1986 年至 2003 年期间每年提高 52%，在 2003 年至 2011 年期间每年提高 23%，但在 2011 年至 2018 年期间放缓至每年仅 7%。尽管延缓登纳德缩放定律的办法显而易见是多核心方案，但这对

控制功耗的帮助有限。

2、经通胀调整的 IT 设备在 1995-1999 年间，价格下降速度加快至每年 23%，随后 2010-2013 年间，价格放缓至每年 2%。

3、20 世纪 90 年代末，CPU 价格下降达到每年 60%（每 9 个月减半），而之前和之后的几年里，改善速度通常为 30%（每两年减半），尤其是笔记本电脑 CPU，2004-2010 年间每年改善 25-35%，2010-2013 年间则放缓至每年 15-25%。

面对这些挑战，半导体行业正在探索多种途径以延续摩尔定律，包括多核心技术、新材料、新架构、以及优化设计流程等。

其中多核心是一个显而易见的可操作方案。英特尔于 2005 年 4 月发布了全球首款双核心处理器 Pentium Extreme Edition 840。此后又将核心数又扩展到了 4 核心、8 核心...到了 2024 年 6 月，至强 6700E 处理器已经达到了 144 核心。但尽管如此，CPU 的核心数量的增加，依然非常有限。

于是技术路线在不同的应用场景下，开始选择 CPU 的替代方案——GPU。

AlphaGo 横空出世

1997 年，IBM“深蓝”击败国际象棋世界冠军加里·卡斯帕罗夫。但近 20 年里，使用人工智能技术的最强围棋程序也只达到业余五段水平，并且仍然无法在不让子的情况下击败专业围棋选手。2012 年，运行在四台 PC 集群上的日本围棋软件 Zen（中文翻译：禅）在五子和四子让子比赛中（人类选手让子）两次击败竹宫正树。2013 年，法国围棋软件 Crazy Stone 在四子让子比赛中击败了石田芳雄。在 AlphaGo 及其开发团队 DeepMind 出现之前，几乎所有研究者都认为在十年内人工智能战胜围棋大师的机会是渺茫的。

伦敦的 DeepMind 公司在 2010 年成立，2014 年被谷歌收购。AlphaGo 研究项目于 2014 年左右启动，旨在测试使用深度学习的神经网络在围棋方面的表现。它比之前的围棋程序有了显著的进步:在与其他可用围棋程序(包括 Zen 和 Crazy Stone)的 500 场比赛中，在单台计算机上运行的 AlphaGo 赢了 499 场比赛。在类似的对决中，在多台计算机上运行的 AlphaGo 赢得了与其他围棋程序进行的全部 500 场比赛，在与在单台计算机上运行的 AlphaGo 进行的比赛中，赢了 77%的比赛。

2015 年 10 月，分布式版本的 AlphaGo 以 5 比 0 击败了欧洲围棋冠军二段职业棋手樊麾。这是计算机围棋程序首次在全尺寸棋盘上击败人类职业棋手。该版本使用了 1202 个 CPU 和 176 个 GPU。

2016 年 3 月，AlphaGo 在首尔与韩国九段职业围棋选手李世石进行了五场比赛，AlphaGo 赢了四场，李世石赢了第四场。AlphaGo 在谷歌的云计算上运行，其服务器位于美国。据《经济学人》报道，它使用了 1920 个 CPU 和 280 个 GPU。从下表可以清晰的看出，GPU 的数量越多，ELO 得分也就越高。

表2：早期的 AlphaGo 配置

配置	线程数	CPU 数量	GPU 数量	Elo 评级
单机	40	48	1	2,181
单机	40	48	2	2,738
单机	40	48	4	2,850
单机	40	48	8	2,890
分布式	12	428	64	2,937
分布式	24	764	112	3,079
分布式	40	1,202	176	3,140
分布式	64	1,920	280	3,168

资料来源：维基百科，国信证券经济研究所整理

2016 年 5 月，谷歌推出了自己的专有硬件 TPU（张量处理单元，Tensor Processing Unit），TPU 是谷歌为神经网络机器学习开发的专用集成电路，使用了谷歌自己的 TensorFlow 软件。与 GPU 相比，TPU 专为大量低精度计算而设计，每焦耳的输入/输出操作更多，而无需用于光栅化/纹理映射的硬件。不同类型的处理器适用于不同类型的机器学习模型。TPU 非常适合 CNN（卷积神经网络），而 GPU 对一些全连接神经网络有优势，CPU 对 RNN（循环神经网络）有优势。

下表可以看到，采用了 TPU 的 AlphaGo Master，只用了 2 块 TPU（单块功耗为 280W），其 ELO 得分就达到了 4858 分，且在未来围棋峰会上以 60：0 的成绩战胜了职业选手。AlphaGo Zero 则是用了 3 块 TPU，对战 AlphaGo（李世石版）成绩达到了惊人的 100:0！而 2017 年 12 月的 AlphaZero 版本，由 4 块 TPU 组成，其 ELO 得分不如 AlphaGo Zero，但成绩却略胜一筹（60:40）。

表3：AlphaGo 的演进

版本	硬件	Elo 评级	日期	结果
AlphaGo（对战范晖）	176 个 GPU，分布式	3,144	2015 年 10 月	5:0 战胜范晖
AlphaGo（对战李世石）	48 个 TPU，分布式	3,739	2016 年 3 月	4:1 战胜李世石
AlphaGo Master	2 个 TPU，单机	4,858	2017 年 5 月	60:0 战胜职业选手，未来围棋峰会
AlphaGo Zero	3 个 TPU，单机	5,185	2017 年 10 月	100:0 战胜 AlphaGo（李世石），对阵 AlphaGo Master 89:11
AlphaZero	4 个 TPU，单机	5,018	2017 年 12 月	60:40 击败 AlphaGo Zero

资料来源：维基百科，国信证券经济研究所整理

在 2018 年之后，Deepmind 就没有再公布在围棋方向上的消息。但 AlphaGo 如同旋风般的出场，给了人工智能在围棋方向上的启发却是深远的：

1、首先，在经历了长时间的沉寂，神经网络终于担负起让机器学习更进一步的担子；

2、并行运算是 AI 的未来：AlphaGo 各种版本进化的历史，也是 GPU 或者 TPU 的贡献较 CPU 更大，并行运算为机器学习更深次探索提供了更为有利的条件；

3、硬件优化的可行性：GPU 是通用型处理器，虽然早期的 AlphaGo 版本使用了多个 GPU 后很快实现了强大的算力，但其功耗也是巨大的。按照单块 CPU 100W，GPU 200W 的功耗计算，2016 年版本的 AlphaGo 使用了 1920 个 CPU 和 280 个 GPU，每下一盘棋，成本就会超过 3000 美元（主要是电费）！而通过在设计，仅 2-4 块 TPU 就能完成类似的，甚至更好的成绩。

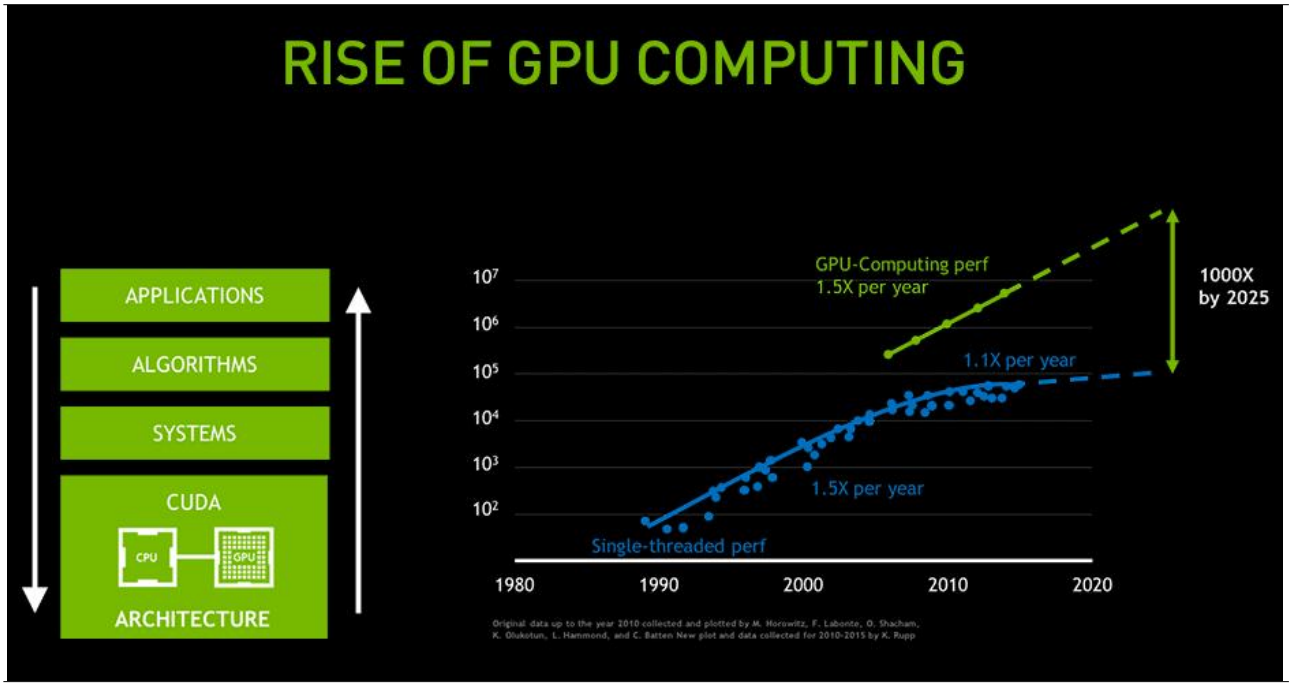
算力指数级进步：GPU 接过接力棒

英伟达创始人兼 CEO 黄仁勋也多次提及摩尔定律已死，并提出黄氏定律（Huang's

law）：GPU 的性能每两年将翻一番以上。他的证据是，2006 年，英伟达 GPU 性能比其他 CPU 高出 4 倍，到了 2018 年英伟达 GPU 比同类 CPU 快 20 倍：即 GPU 每年快 1.7 倍。

实际上，黄仁勋的预测也经常改变。下图是 2017 年黄仁勋提及的未来英伟达 GPU 的速度是每年提升 50%（1.5X）。也有人根据英伟达 AI 性能 10 年提升 1000 倍，认为黄氏定律应该是每年翻倍，即 10 年正好翻 2^{10}即 1000 倍左右。

图4：英伟达对 GPU 速度预测（2017）



资料来源：英伟达，国信证券经济研究所整理

2023 年 9 月，英伟达首席科学家比尔·戴利(Bill Dally)提及，英伟达单个 GPU 在 AI 推理方面的性能大幅提升了 1000 倍，戴利还将这 1000 倍分成了 4 部分（1000=16×12.5×2.5×2），其中：

1、算法优化 16 倍（finding simpler ways to represent the numbers computers use to make their calculations）；

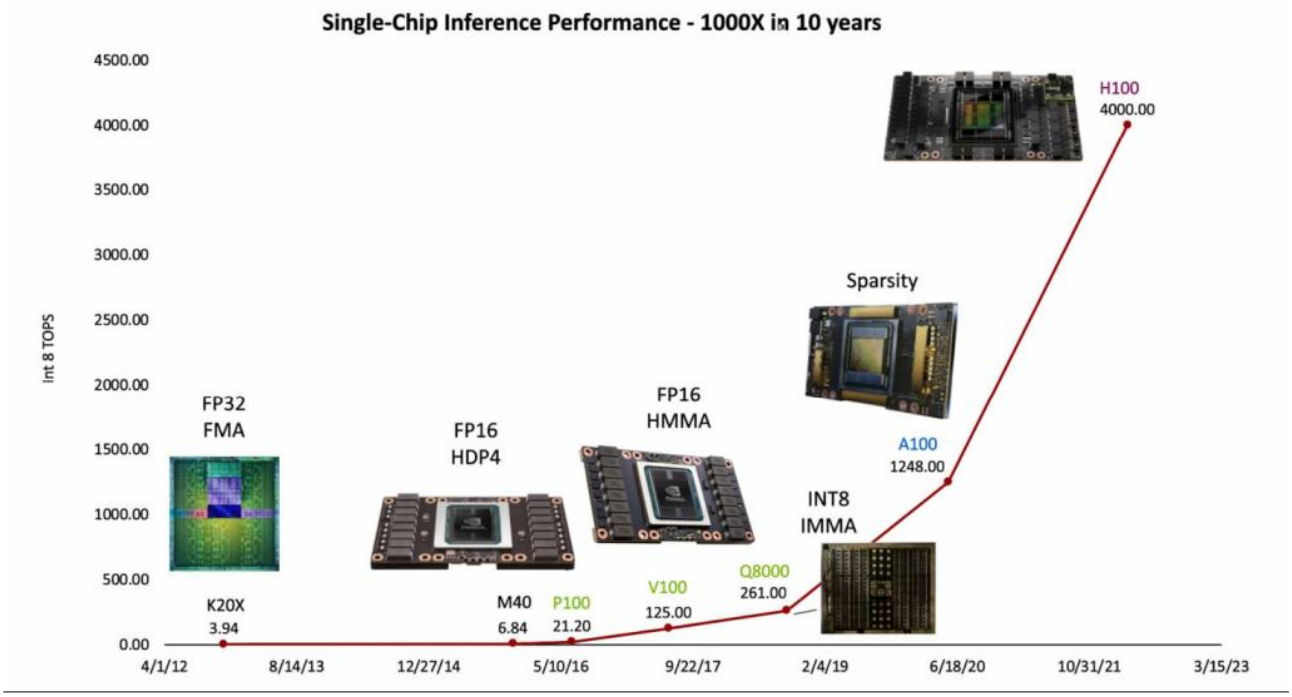
2、指令优化 12.5 倍（crafting advanced instructions that tell the GPU how to organize its work）；

3、结构优化 2 倍（structural sparsity）；

4、制程进步 2.5 倍（从 28nm 迁移到 5nm）。

可见，从这个案例来看，英伟达的性能提升，主体并非制程进步，而是算法与指令的优化。

图5：英伟达 GPU 在 10 年的时间里，AI 推理速度提升了 1000 倍



资料来源：英伟达，国信证券经济研究所整理

记者乔尔·赫鲁斯卡于 2020 年撰文称，“黄氏定律根本不存在”，称其为一种建立在摩尔定律带来的收益之上的“幻觉”，现在断定该定律是否存在还为时过早。非营利性研究机构 Epoch 发现，2006 年至 2021 年间，GPU 性价比（以 FLOPS/$为单位）每 2.5 年翻一番，比黄氏定律预测的要慢得多。一个简单的现象是，尽管我们讨论英伟达的单个 GPU 性能提升是不假，但是其价格确实水涨船高，例如大模型出现之后，2023 年的英伟达的 H100 最高炒到 3-4 万美元一块，这显然比曾经的价格要高出很多。

但无论是每年 2 倍，1.5 倍，还是每两年 1 倍，GPU 的确接过了 CPU 的接力棒，更有效率地推高了计算能力，这一点是毋庸置疑的。

早期的 GPU 是图像处理设备，或者叫显卡。1995 年 11 月，3DFX 公司的 Voodoo 显卡问世，Voodoo 是当时市场占有率最高的 3D 显卡，市场份额高达 85%。相较而言，英伟达在当时诸多显卡公司中不算显眼，但得益于当时英伟达快速追随了微软，发布了适配微软公司的 Direct3D 7 标准的 Geforce256 显卡，使得英伟达迅速成为显卡市场的佼佼者。

图6：3DFX 公司的 Voodoo 显卡（1995 年）

资料来源：百度百科，国信证券经济研究所整理

图7：英伟达公司的 Geforce256（1999 年）

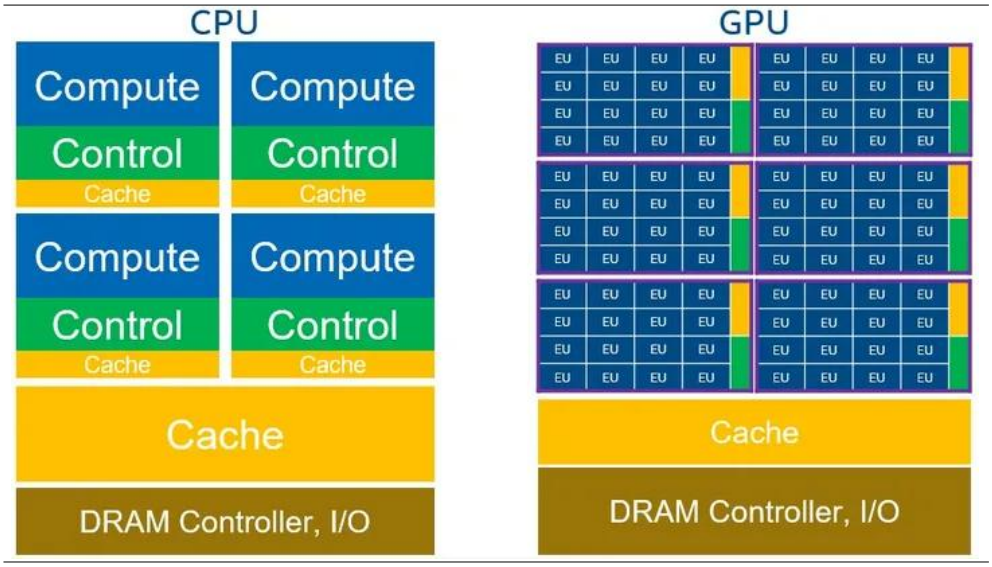
资料来源：百度百科，国信证券经济研究所整理

2000 年，英伟达收购了 3DFX；而 2006 年，AMD 收购了 ATI。似乎行业就这样趋于平静，但很快技术上的变化打破了这种平静。

2003 年，人们发现了基于 GPU 来解决一般线性代数问题，其运行速度比在 CPU 上更快。这些早期将 GPU 用作通用处理器的工作，需要根据图形重新表述计算问题，正如图形处理器的两个主要 API，OpenGL 和 DirectX 所支持的那样。于是人们思考，是否能够建立一种通用编程语言和 API 来减少其中的繁琐与转换。

这就出现了 GPGPU 概念（General-purpose GPU）。GPGPU 的通用计算是使用图形处理单元(GPU)来执行传统上由中央处理器(CPU)处理的应用程序中的计算，而图形处理单元(GPU) 通常仅作计算机图形处理。在一台计算机中使用多个视频卡或大量图形芯片，使图形处理本来就并行的特性进一步并行化。本质上，GPGPU 是一个或多个 GPU 与 CPU 之间的一种并行处理，可将数据视为图像或其他图形形式进行分析。虽然 GPU 以较低的频率运行，但它们的核心数量通常要多很多倍。因此，GPU 每秒可以处理的图片和图形数据比传统 CPU 多得多。将数据迁移到图形形式，然后使用 GPU 对其进行扫描和分析可以大大提高速度。

图8：CPU 与 GPU 的架构区别



资料来源：知乎，国信证券经济研究所整理

英伟达认为，GPGPU 对于显卡公司来说，是个非常有前景的发展方向。在 2006 年，英伟达发布了 CUDA 架构（Compute Unified Device Architecture，统一计算设备架构）。CUDA 是一个专有的并行计算平台和应用程序编程接口(API)，允许软件使用某些类型的图形处理单元(GPU)进行加速通用处理。CUDA API 是 C 语言的扩展，它增加了在 C 中指定线程级并行性的能力，以及指定 GPU 设备特定的操作。CUDA 是一个软件层，它可直接访问 GPU 的虚拟指令集和并行计算元素，以执行计算内核。除了驱动程序和运行时内核之外，CUDA 平台还包括编译器、库和开发人员工具，以帮助程序员加速他们的应用程序。CUDA 旨在与 C、C++、Fortran 和 Python 等编程语言配合使用。这种可访问性使并行编程专家可以更轻松地使用 GPU 资源，而之前的 Direct3D 和 OpenGL 等 API 则需要高级图形编程技能。基于 CUDA 的 GPU 还支持 OpenMP、OpenACC 和 OpenCL 等编程框架。

最初的 CUDA SDK 于 2007 年 2 月面向 Microsoft Windows 和 Linux 公开。后来在 2.0 版中添加了 Mac OS X 支持。CUDA 适用于 G8x 系列及以后的所有英伟达 GPU，包括 GeForce、Quadro 和 Tesla 系列。

此后，英伟达显卡中都包含了支持 CUDA 运算的核心。而这个核心数量，要比 CPU 的核心数量多得多。换句话说，在 CPU 无法通过先进制程去提速，而通过核心的扩展来实现提速这条路走的不顺畅时，GPU 通过扩展核心来提速却容易得多。

下表列示了英伟达部分 GPU 与 CUDA 核心数，可以看出，从 2008 年以来，其核心数从 240 个增加到了 2022 年的 18432 个。

表4：部分英伟达 GPU 的 CUDA 核心数

GPU 型号	发布时间	制程工艺（纳米）	CUDA 核心数
GeForce 8800 GTX	2006 年 11 月	90	不适用
GeForce GTX 280	2008 年 6 月	65	240
GeForce GTX 480	2010 年 4 月	40	480
GeForce GTX 580	2010 年 11 月	40	512
GeForce GTX 680	2012 年 3 月	28	1536
GeForce GTX 780	2013 年 5 月	28	2304
GeForce GTX 980	2014 年 9 月	28	2048
GeForce GTX 1080	2016 年 5 月	16	2560
GeForce RTX 2080 Ti	2018 年 9 月	12	4352
GeForce RTX 3090	2020 年 9 月	8	10496
H100	2022 年 3 月	4	18432

资料来源：英伟达，国信证券经济研究所整理

另外一个角度，英伟达 GPU 经历了多年的演进，已经先后经历了 10 个主要架构（micro architectures），从 2006 年的 Tesla 架构，到 2024 年的 Blackwell 架构，即大约每 2 年就会有一个新的架构诞生。每一次架构的迭代，大都伴随制程工艺的进步，速度的提升，以及功能的改善（例如 Volta 架构引入 Tensor 核心、Turing 架构引入 RT 核心）。

但值得注意的是，我们列举了重点型号 GPU 的功率，发现其并非保持不变的，而是大约在 18 年的时间里提升了 4-5 倍。例如，Telsa 架构下的 GeForce 8800 GTX 功率仅为 175W；到了 Volta 架构的 Tesla V100，功率到了 300W；而在 Hopper 架构下的 H100，功率达到了 700W。这也说明，英伟达比英特尔在扩展核心上技高一筹这并不假，或者说，GPU 接替 CPU，担负起推动人类算力革命的使命。

但此间也有个问题，即英伟达并未在算力扩张的情况下保持功率不变，这与我们理解的“摩尔定律”是不相符的：这好比摩尔定律定义的是一块 CPU 每 18 个月速度翻倍，而英伟达实现过程更像是把多块芯片做到一起（因为 GPU 核心可以很多）

而宣称是一块芯片。从这个角度说，“黄氏定律”不宜称之为一个定律，只能代表英伟达追求速度的一种标榜。

表5：英伟达通用 GPU 的架构

架构名称	发布时间	制程工艺	代表型号	功率范围	说明
Tesla	2006 年	90nm	GeForce 8800 GTX	175W	引入统一着色器架构
Tesla	2008 年	65nm/55nm	GeForce GTX 280	236W	增强了 CUDA 核心
Fermi	2010 年	40nm	GeForce GTX 480	250W	第一个支持 DirectX 11 的架构
Kepler	2012 年	28nm	GeForce GTX 680	195W	提升了能效比
Maxwell	2014 年	28nm	GeForce GTX 980 Ti	250W	优化内存带宽
Pascal	2016 年	16nm	GeForce GTX 1080 Ti	250W	高性能与能效
Volta	2017 年	12nm	Tesla V100	300W	面向数据中心，首次引入 Tensor Core，专注于深度学习和 AI 应用
Turing	2018 年	12nm	GeForce RTX 2080 Ti	250W	引入了 RT Core，支持实时光线追踪
Ampere	2020 年	8nm	GeForce RTX 3090	350W	支持 RTX IO
Hopper	2022 年	4nm	H100	700W	专为 AI 和数据中心设计
Blackwell	2024 年 3 月	3nm	GB200	1000W	专为 AI 和 HPC 设计，支持 HBM3E 显存，具有高带宽和低功耗特性。

资料来源：英伟达，国信证券经济研究所整理

观察近百年来计算机发展的历史，可以得到计算机性能（FLOPS）与成本的关系。下表列示了提供最低每 GFLOPS 成本的平台，对应的成本（换算成 2022 年的美元计价）。可以看出，为了推动成本的下降，从 2010 年之后，表中开始频繁的出现了 GPU 以替代 CPU，尤其是 2020 年之后，表中的方案都是以 GPU 实现的。

例如索尼 PlayStation 4 采用的是 AMD 美洲豹 CPU 但集成了 GPU，PlayStation 5 采用了定制版 RDNA 2 架构的 AMD GPU，Xbox Series X 包含了一个具备 12TFLOPS 算力的 GPU，RTX 4090、镭龙 RX 7600 都是 GPU。

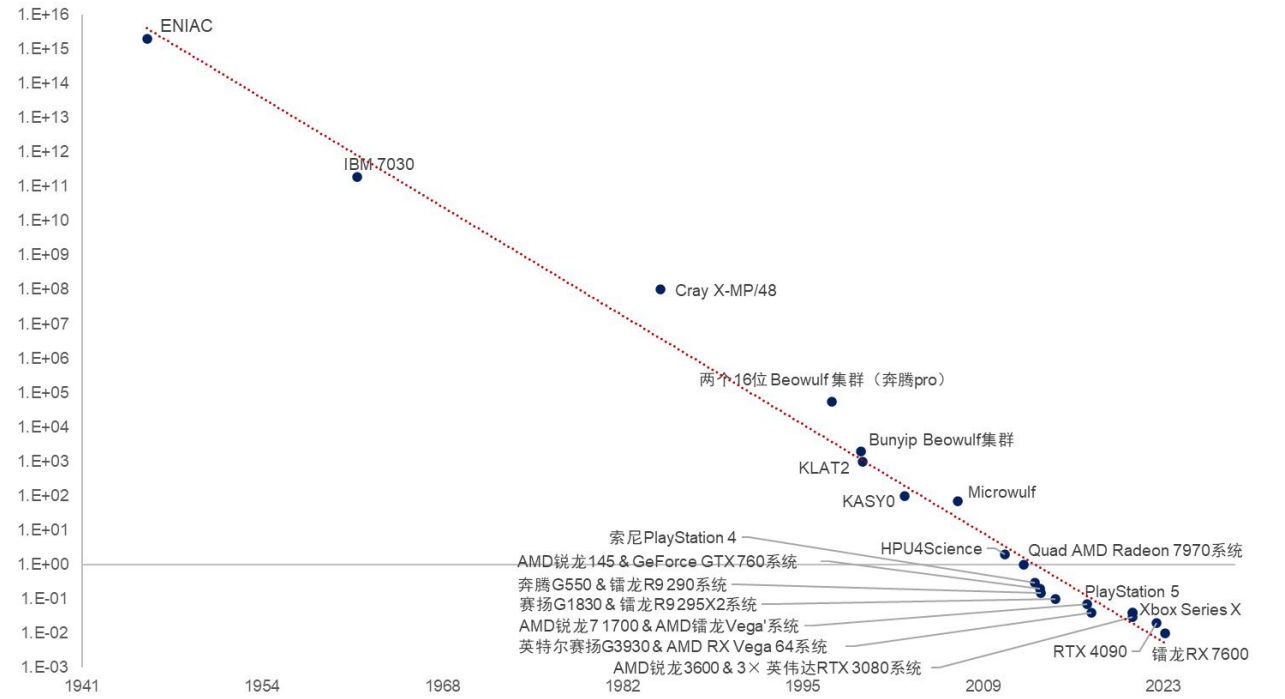
表6：每 GFLOPS 成本变化

日期	每 GFLOPS 成本（美元）		提供最低每 GFLOPS 成本的平台
	原始成本	调整成本（2022）	
1945 年	130 万亿美元	2E+15	ENIAC
1961 年	200 亿美元	1.96E+11	IBM 7030
1984 年	$20,000,000	$100,000,000	Cray X-MP/48
1997 年	$30,000	$55,000	两个 16 位 Beowulf 集群（奔腾 pro）
2000 年 4 月	$1,000	$2,000	Bunyip Beowulf 集群
2000 年 5 月	$640	$1,000	KLAT2
2003 年 8 月	$90	$100	KASY0
2007 年 8 月	$50	$70	Microwulf
2011 年 3 月	$1.80	$2	HPU4Science
2012 年 8 月	$0.75	$1	Quad AMD Radeon 7970 系统
2013 年 6 月	$0.22	$0.30	索尼 PlayStation 4
2013 年 11 月	$0.16	$0.20	AMD 锐龙 145 & GeForce GTX 760 系统
2013 年 12 月	$0.12	$0.15	奔腾 G550 & 镭龙 R9 290 系统
2015 年 1 月	$0.08	$0.10	赛扬 G1830 & 镭龙 R9 295X2 系统
2017 年 6 月	$0.06	$0.07	AMD 锐龙 7 1700 & AMD 镭龙 Vega 系统
2017 年 10 月	$0.03	$0.04	英特尔赛扬 G3930 & AMD RX Vega 64 系统
2020 年 11 月	$0.03	$0.03	AMD 锐龙 3600 & 3× 英伟达 RTX 3080 系统
2020 年 11 月	$0.04	$0.04	PlayStation 5
2020 年 11 月	$0.04	$0.04	Xbox Series X
2022 年 9 月	$0.02	$0.02	RTX 4090
2023 年 5 月	$0.01	$0.01	镭龙 RX 7600

资料来源：维基百科，国信证券经济研究所整理

我们将调整后的成本（2022）与时间线绘制成图，可以清晰地观察出，它呈现的是一个指数曲线（图中为对数坐标系）。

图9：不同计算设备提供的每 GFLOPS 成本



资料来源：维基百科，国信证券经济研究所整理

我们将调整后的成本（2022 年）与时间线绘制成图，可以清晰地观察出，它呈现的是一个指数曲线（图中为对数坐标系）。

为了观察每 GFLOPS 成本的变化速度，我们再将时间分阶段来观察：

1、从 1945 年以来（拥有数据的最长历史），77 年的时间里，单位算力（GFLOPS）的年复合成本下降幅度为 40.24%；

2、1961 年以来的 61 年里，1984 年以来的 38 年里，1997 年以来的 25 年里，2007 年以来的 16 年里，年复合成本下降幅度介于 39%-45%之间；说明技术推动相对顺利；

3、从 2011 年以来的 12 年里，年复合成本下降幅度为 35.28%，以及 2015 年以来的 8.3 年里，年复合成本下降幅度仅为 24.14%，这说明随着摩尔定律遇到极限挑战之后，技术推动效果也在明显放缓。

表7：不同时间周期下每 GFLOPS 的复合成本降幅

	时间（年）	年复合成本下降幅度
1945 年以来	77.4	(40.24%)
1961 年以来	61.4	(39.27%)
1984 年以来	38.4	(45.14%)
1997 年以来	25.3	(45.79%)
2007 年以来	15.8	(42.98%)
2011 年以来	12.2	(35.28%)
2015 年以来	8.3	(24.14%)

资料来源：维基百科，国信证券经济研究所整理

这带来了一些深远的影响：

1、更大的资本投入：想要获得更大的算力，如果采用“时不我待”的态度，则需

支付更高的成本；

2、体积不能更小，但能更大：由于 GPU 单卡功率变大，因此体积无法更小，例如在 5 年之内，目前似乎尚无法看到一个手持设备的算力可以赶上主流 GPU 的水平，则应用的方向朝着体积更大去演进，如云计算可以忽略设备的占地，如汽车也拥有相对较大的空间，如台式机相对可以配置较好的 GPU 显卡，而笔记本、手机、智能穿戴等设备，则短期较难享受到 GPU 算力革命的巨大成果；

3、有利于中国的追赶：由于年复合成本下降速度放缓，换句话说，先发者对于后来者的比较优势也在缩小，这给中国芯片业迎来宝贵的时间窗；

4、等待着新技术的突破：在技术瓶颈期，往往新技术才能冲破。例如硅半导体的可能替代品：碳化硅、石墨烯、金刚石、其他 III-V 族化合物（如砷化镓 GaAs）、II-VI 族化合物（如硫化镉 CdS）等，或者量子计算等更前沿的技术。

应用的助力：比特币、云计算、新能源汽车

有三大应用场景对 GPU 的发展起到了重要的推动作用。它们分别是加密货币、云计算、新能源汽车。

加密币的诞生

1、比特币拉动了全网算力的提升

2008 年 10 月，中本聪撰写的白皮书《比特币：一种点对点电子现金系统》（A Peer-to-Peer Electronic Cash System）问世了。比特币软件作为开源代码则是在 2009 年 1 月发布。2009 年 1 月 3 日，中本聪挖出了比特币链的起始区块（即创世区块），比特币网络由此诞生。2010 年 5 月 22 日，已知的第一笔比特币商业交易发生在程序员拉斯洛·汉耶茨以 10,000 比特币购买了两个帕帕约翰披萨时，这一天后来被称为“比特币披萨日”。

比特币从诞生之日起，就伴随着旷日持久的争论与非议。欧洲央行认为，比特币提供的货币去中心化理论根源于奥地利经济学派，尤其是哈耶克的《货币非国家化》（The Denationalization of Money）一书，他在书中主张在货币的生产、分配和管理方面建立完全的自由市场，以结束中央银行的垄断。《比特币独立宣言》认为比特币意识形态的本质是将货币从社会和政府控制中解放出来。《经济学人》将比特币描述为“一个技术无政府主义项目，旨在创建现金的在线版本，让人们可以进行交易而不受恶意政府或银行干扰”。这些哲学思想最初吸引了自由主义者和无政府主义者。经济学家保罗·克鲁格曼认为，只有银行怀疑论者和犯罪分子才会使用比特币等加密货币。

不少经济学家、投资者都将比特币描述为潜在的庞氏骗局。巴菲特也持有类似的观点，他多次在公开场合表达对比特币的批评，他认为比特币没有生产能力，其价值完全依赖于市场投机和需求，而不是基于任何实际的生产或服务。他曾经比喻比特币为“老鼠药”，并指出比特币的价值波动巨大，缺乏稳定的现金流，这使得比特币难以被视为一种真正的资产。但法律学者埃里克·波斯纳（Eric Posner）不同意这种观点，因为“真正的庞氏骗局需要欺诈；相比之下，比特币看起来更像是一种集体妄想”。2014 年世界银行的一份报告也得出结论，比特币不是故意的庞氏骗局。

图10：比特币的价格与市值

资料来源：oklink.com，国信证券经济研究所整理

由于比特币的产生机制是：

1、工作量证明：使用工作量证明（PoW,Proof of Work）机制来确定哪些节点有权将新的交易记录添加到区块链上。节点通过解决一个极其复杂的数学难题来达成这一目标，这个难题涉及到对新区块的头信息进行哈希运算，整个过程需要大量的计算尝试，因此需要消耗大量的计算资源；

2、网络难度：比特币网络通过调整挖矿难度来维持平均每 10 分钟产生一个新区块的速度。如果这段时间内区块生成速度过快，则难度会上升；如果过慢，则难度会下降。这样可以确保即使计算能力发生变化，区块生成速率也能保持相对稳定；

3、四年减半：成功挖出新区块的矿工将获得一定数量的比特币作为奖励，大约每四年减半一次。

这三个机制导致了先投入先得利，高算力高回报的竞争结果。因此它不可避免的粗发了显卡（GPU）的抢购潮与军备竞赛。

例如，2010 年 7 月，比特币的全网平均算力为 172M h/s，2024 年的 12 月（14 年以后）比特币的全网平均算力为 751E h/s（E=1000P，P=1000T，T=1000G，G=1000M），也就是说，14 年全网算力增加了 4 万亿倍（$4.4*10^{12}$），每年复合增速 6.4 倍！或者说每 4.1 个月全网算力翻番。当然，由于早期的全网算力小的可怜，既没有人关注，也多用电脑（CPU）来挖矿，等到后期关注度较高且开始大量使用 GPU 之后，算力的上涨速度就降下来了。

图11：比特币的全网平均算力

资料来源：oklink.com，国信证券经济研究所整理
比特币全网平均算力，即全网 BTC 矿工平均每秒可以完成的哈希计算的数量

比特币价格几次大涨：

2011 年：比特币的价格首次突破了 10 美元的大关，并在 6 月份迅速上涨至 30 美元左右，之后迅速回落。这是比特币早期的一次显著增长；

2013 年：从 2012 年年底 10 元左右上涨到在 2013 年 12 月，比特币价格从 10 美元飙升至每枚 1200 美元左右。这次增长主要是由于市场对比特币作为一种新兴金融资产的兴趣增加；

2017 年：比特币在 2017 年经历了非常显著的增长，价格从年初的约 1000 美元左右一路上涨，到了年底达到了近 20000 美元的历史高点。这次大涨的原因包括市

场兴趣激增、更多的机构投资者进入市场以及全球经济不确定性的增加；

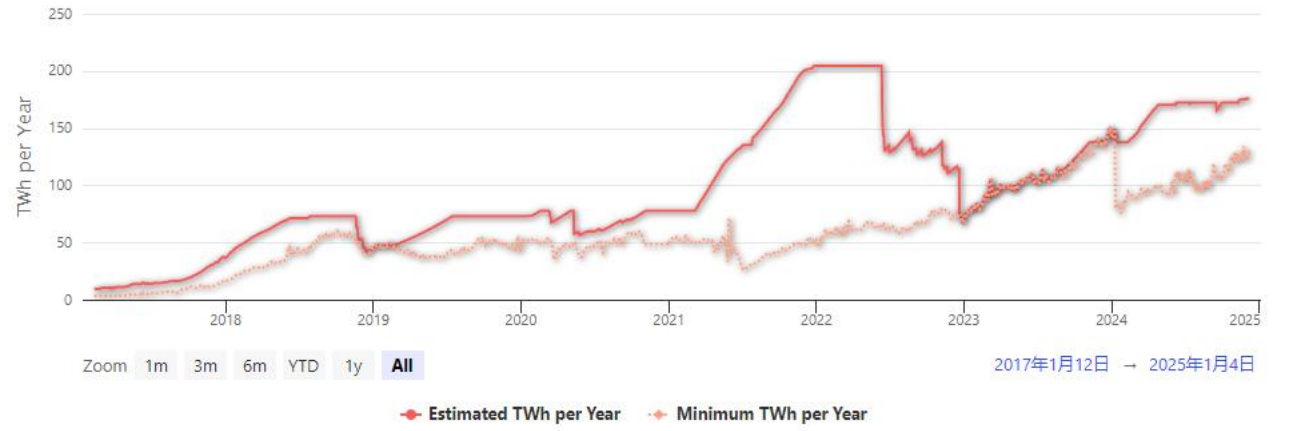
2020-2021 年：受到新冠疫情的影响，美联储大放水，比特币被视为一种避险资产，吸引了大量投资。一些大型机构投资者和上市公司开始持有比特币作为资产储备，进一步推高了价格。2021 年初，比特币价格突破了 5 万美元，并在接下来几个月内继续攀升，最高达到 6 万美元以上；

2023-2024 年：全球央行开始增持黄金，比特币作为另类资产，也开始受到追捧。随着比特币基金的成立以及美国宣布要将比特币作为战略储备，价格从 2022 年底的 14000 美元来到了 2024 年底的 10 万美元。

比特币在每次大涨中，都会极大刺激对显卡（GPU）的需求。例如以 17 年矿潮为例，2 个月挖矿（并除去电费）就可回本，年化收益率高达 600%。

随着算力猛涨，到了 2022 年，全球比特币的矿机高峰耗电量达到了 204TWh。

图12：比特币挖矿的耗电量



资料来源：digiconomist.net，国信证券经济研究所整理

而 2022 年，全球总用电量为 24398Twh。也就是说，巅峰期矿机用电量接近全球用电量的 0.8%，这大约是全球用电量排名第 20 名左右国家的水平，也有研究表明，矿机的平均用电量（而非峰值）占全球用电量的 0.4-0.6%。

2017 年 12 月，芝加哥商品交易所(CME)推出了首个比特币期货；2021 年 10 月，ProShares 的首只比特币 ETF，BITO 在芝加哥商品交易所上市；2024 年 1 月，11 只美国现货比特币 ETF 开始交易，首次在美国证券交易所提供对比特币的直接投资。其中规模最大的 ETF 是贝莱德管理的 iShares 比特币信托（IBIT），2024 年上半年流入约 200 亿美元。

截至 2023 年 6 月，River Financial 估计比特币拥有 8170 万用户，约占全球人口的 1%，但在现货 ETF 发行之后，比特币流动性将明显改善，成为不可忽视的另类投资品种。

表8：全球主要国家/地区的用电量

2022 年	国家/地区	用电量（TWh）	人口，百万人均用电量（兆瓦时）	
	世界	24,398	7,960	3.1
1	中国	7,214	1,443	5.0
2	美国	4,272	336	12.7
3	印度	1,403	1,401	1.0
4	日本	1,132	126	9.0
5	俄罗斯	934	146	6.4
6	加拿大	595	38.1	15.6
7	韩国	553	51.2	10.8
8	巴西	550	215	2.6
9	德国	539	82.2	6.6
10	法国	463	67.7	6.8
11	沙特阿拉伯	317	36	8.8
12	英国	312	68.4	4.6
13	印度尼西亚	308	276	1.2
14	意大利	300	60	5.0
15	墨西哥	296	127	2.3
16	伊朗	280	83.3	3.4
17	火鸡	264	84	3.1
18	中国台湾	257	23.8	10.8
19	西班牙	246	46.8	5.3
20	南非	233	60	3.9

资料来源：维基百科，国信证券经济研究所整理

2、非 PoW 机制提高了记账效率

比特币是去中心化的产物，那么去中心化也带来的效率上的妥协。一个中心化的网络交易速度可以很快（试想阿里巴巴在 2020 年双十一的交易量可达每秒 58.3 万笔）。而在比特币的网络里每秒可以处理的交易量非常有限，通常在每秒 3-7 笔交易左右。这是因为每个区块只能容纳有限的交易，并且区块的生成时间固定为 10 分钟。为了达到较高的安全性，通常建议等待 6 个区块的确认时间，即大约 1 小时。于是，比特币如果与黄金类比，可以免去实物搬运的麻烦；但和现代货币的快捷支付相比，这又成了它最大的掣肘。

以太坊的诞生主要是弥补比特币在交易功能上的不足。以太坊白皮书于 2013 年发布，公链于 2015 年 7 月启动。先以与比特币共同的工作量证明（PoW）的算法来增加可信度，再逐步转换成权益证明（PoS）以增加效率。经历了 2015 年的“边境”、2016 年的“家园”、2017 年的“都会”三个版本后，以太坊迎来了 2020 年的“宁静”版本。2022 年 9 月，以太坊合并完成，主网与 PoS 共识层信标链（Beacon 链）结合、将此前 PoW 工作量证明机制转变为 PoS 权益证明机制，宣布以太坊正式进入 2.0 时代。

以太坊最重要的技术贡献是智能合约。智能合约是存储在区块链上的程序，可以协助和验证合约的谈判和执行。纽约时报称以太坊平台是一台公共电脑，由众多用户构成的网络来运转，通过以太币来分配和支付这台电脑的使用权。经济学人则说明智能合约可以让众多组织的数据库得以用低廉的成本交互，并且让用户写下精密的合约，功能之一是产生去中心化自治组织，也就是一间只是由以太坊合约构成的虚拟公司。

以太坊 2.0 通过采用 PoS 共识机制、引入分片技术和优化技术等手段，使得网络能够处理更多的交易，并且以更快的速度完成这些交易。这些改进对于提高以太坊网络的整体性能至关重要，尤其是在面对日益增长的去中心化应用需求时。

甚至在 NFT 应用场景中，以太坊的记账速度还显得不够高效。例如 Solana 的出现就是希望自己成为更快的记账网络，Solana 由前高通、英特尔及 Dropbox 的工程师团队于 2017 年末创立，2020 年 3 月网络上线。它采用了历史证明（PoH）与权益证明（PoS）混合共识机制，理论上的交易处理速度可以达到每秒数十万笔交易（TPS），目前市值排名第六。

目前全球市值前五大区块链网络分别比特币、以太坊、泰达币（稳定币）、币安币、XRB（瑞波币）。除了比特币与早期的以太坊，其他网络都没有以工作量证明（PoW）为共识机制。

图13：全球市值排名前 5 的区块链网络

Rank	Name	Price	Market Cap	VWAP (24Hr)	Supply	Volume (24Hr)	Change (24Hr)
1	Bitcoin BTC	$99,721.83	$1.97t	$99,814.40	19.79m	$8.91b	0.30%
2	Ethereum ETH	$3,981.19	$479.56b	$3,999.25	120.44m	$6.84b	-0.20%
3	Tether USDT	$1.00	$138.01b	$1.00	137.93b	$29.28b	-0.08%
4	BNB BNB	$747.71	$124.75b	$749.46	166.80m	$510.47m	0.51%
5	XRP XRP	$2.56	$116.15b	$2.51	45.40b	$3.79b	4.71%

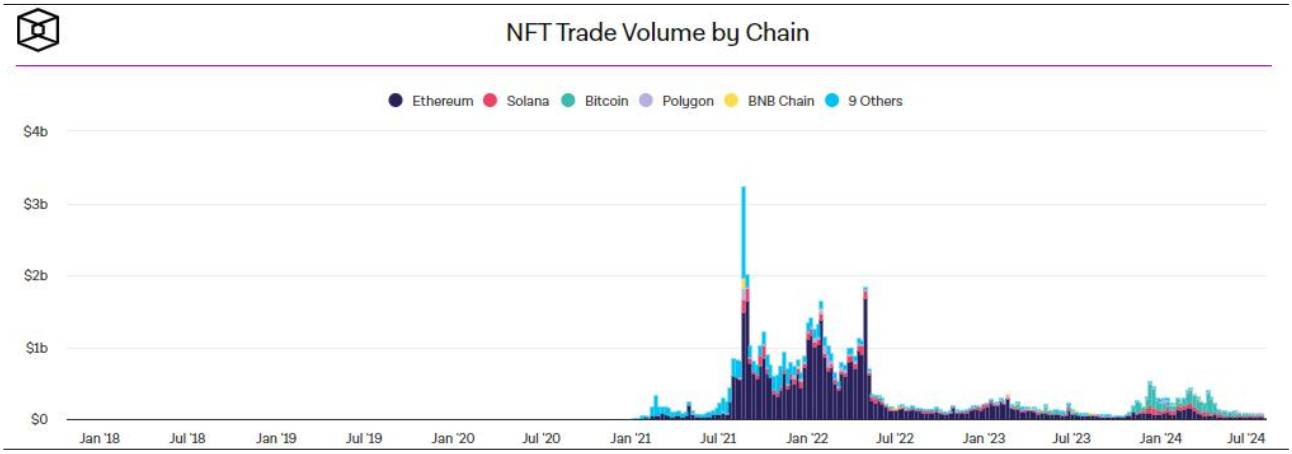
资料来源：coincap.io，国信证券经济研究所整理

3、NFT 与区跨链游戏丰富了加密币的使用场景

随着以太坊、Solana、Polygon 等高效记账区块链的发展，诸多应用的支付开始转移到它们上。比较典型的两大场景是 NFT 区块链游戏。

非同质化代币(NFT，Non-fungible token)是一种记录在区块链上的唯一数字标识符，用于证明所有权和真实性。它不能被复制、替换或细分。NFT 的所有权记录在区块链中，所有者可以转让，从而允许 NFT 出售和交易。

图14：NFT 交易规模



资料来源：theblock.co，国信证券经济研究所整理

NFT 交易额从 2020 年的 8200 万美元增至 2021 年的 170 亿美元，在 2021 年 8 月 21 日的一周内，NFT 单周交易额高达 32.4 亿美元。但在 2022 年，NFT 市场崩溃，交易量大幅下行，目前 NFT 交易额每周不到 1 亿美元，交易主要集中在以太坊、

Solana、比特币、Polygon 网络上。

美国数字艺术家温克尔曼，艺名 Beeple，耗时 14 年，自 2007 年起每天创作一幅画作并上传到网络上，最终把 5000 张作品图拼接成一个 316MB 的 JPG 文件，命名为《每一天：最初的 5000 天》。这件拍品以 100 美元起拍，在互联网上历经长达 15 天的竞标，2021 年 3 月在佳士得拍卖行以 6930 万美元落槌。

买家是新加坡程序员 Sundaresan，他是一名加密货币投资者，也是 Metapurse NFT 项目的创始人。由于买家和卖家都有既得利益，希望推高作品的价格，以引起人们对与其他 20 件 Beeple 作品相关的投机资产的关注并推动其销售，他们称之为“B20 代币”。这些代币的价格在媒体报道 Everydays 拍卖期间达到顶峰，随后暴跌，Sundaresan 持有 59%的权益，Beeple 持有 2%的权益。正因为如此，一些观察人士将这次拍卖描述为一场宣传噱头和骗局。

图15：《每一天：最初的 5000 天》（Everydays: The First 5,000 Days）

资料来源：百家号，国信证券经济研究所整理

批评者将 NFT 市场的结构比作金字塔或庞氏骗局，早期投资者以牺牲后来者为代价去获利。2022 年 6 月，比尔·盖茨表示，NFT 是“100%基于更大傻瓜理论”。

乐观者则将 NFT 市场比作 17 世纪的郁金香狂热，称任何投机泡沫都需要技术进步才能让人们“兴奋”，而这种热情的一部分来自对产品的极端预测（即泡沫可以持续很长时间）。

除了 NFT，区块链游戏也在相似的时间大火。最早使用区块链技术的最著名游戏

之一是 CryptoKitties，由 Axiom Zen 于 2017 年 11 月为个人电脑推出。玩家可以使用以太坊加密货币购买宠物，玩家可以将其与其他宠物繁殖，创造出具有综合特征的后代作为新的宠物。这款游戏在 2017 年 12 月成为头条新闻，当时一只虚拟宠物的售价超过 100,000 美元。当时大约 30%的以太坊交易都是为游戏而做的，拥堵延迟了玩家的交易，也有了后来的以太网的升级。

Sky Mavis 于 2018 年发布的 Axie Infinity 是一款“边玩边赚”（Play to earn）的游戏，游戏通过活动激励玩家购买并改进 NFT，然后由发行商转售给其他玩家，玩家会因其劳动获得报酬。在游戏最受欢迎的菲律宾，一些玩家通过玩游戏赚到了足够的钱来支付生活费用。在 2022 年初的一次黑客攻击之后，Axie Infinity 发行商被盗 6 亿多美元，游戏玩家数量大幅下降，游戏经济受到影响。由于代币价值暴跌，Sky Mavis 在其网站和营销中删除了对“边玩边赚”的提及。

图16：CryptoKitties

资料来源：百度百科，国信证券经济研究所整理

图17：Axie Infinity

资料来源：baiozhuntuixing.com，国信证券经济研究所整理

部分评论家将区块链“边玩边赚”模式描述为金字塔骗局。“边走边赚”游戏（一种以步行为奖励加密货币的健身游戏）的竞争对手开发商争相指责对方是庞氏骗局，同时努力寻找“解决庞氏经济学问题”的方法。

不可否认的是，在区块链游戏火热之时，无一例外是比特币/以太坊价格上涨之时，上涨促使游戏经济繁荣，会带来更多的玩家，因此即便亏钱的运营模式也会因为玩家的陡然增加而变得健康。玩家们大都不会因为游戏的内容而被吸引，而是流连于加密币的上涨所带来的财富增加。

NFT 与游戏并非没有创新，DAO（去中心化自治组织，Decentralized autonomous organization）是其独特的组织形式。发起者可以是个人，可以是公司，但他们在项目中，都成为了 DAO 的一部分。DAO 的组织形式探索类似于人类社会早期在数千年中尝试的多种不同治理方式，因此它绝对不是一件简单的工作，也绝不会有个标准答案，与 DAO 对立的是已经成熟了几百年的公司制。所以这类 DAO 商业组织的确切法律地位通常不明确，并且可能因司法管辖区而异。2021 年 7 月，怀俄明州成为美国第一个承认 DAO 为法人实体的州，关于 DAO 的法律法规完善还有很长一段路要走。

4、元宇宙概念的火热

2021 年称为元宇宙元年。3 月沙盒游戏平台 Roblox 在纽交所成功上市，事件经过各大媒体报道后引发了各界关注，形成“元宇宙”现象。4 月与 10 月元宇宙在 YouTube 上的搜索量达到了两次峰值，在谷歌上于 2021 年 10 月 24 日达到了峰值。2021 年 10 月 28 日，Facebook 创始人扎克伯格宣布公司旗下部分产业更改全新的名字为 Meta，名字来源为元宇宙 MetaVerse 的前缀。

如果回顾时间线，无论是 AR 还是 VR，在这一年里并未有重大的技术突破。而加密币与 NFT 的火热确实发生在 2021 年，尤其是 NFT 的成交量在 2021 年达到峰值。因此说，元宇宙的火热与区跨链、NFT 的相关度显然更高。

1992 年，元宇宙一词诞生于科幻小说《雪崩》。小说中“元宇宙”世界是一个平行于现实世界的虚拟共享世界。在元宇宙中，用户可以通过佩戴便携式终端、眼镜和其他体感设备保持与元宇宙的持续连接。诸多文艺作品诠释的元宇宙可能更丰满，除了《雪崩》之外，这些作品还包括：《黑客帝国》、《玩家 1 号》、《神经唤术士》、《刀剑神域》、《加速世界》、《夏日大作战》、《龙与雀斑公主》、《西部世界》等。

维基百科对元宇宙的描述是：元宇宙主要探讨一个持久化和去中心化的线上三维虚拟环境。此虚拟环境将可以通过 VR、AR、电话、个人电脑和电子游戏机进入人造的虚拟世界。此虚拟世界需要各种科技如区块链、人工智能、扩增实境、机器视觉。Roblox 给出的元宇宙中共包含八大要素，分别为：身份、朋友、沉浸感、低延迟、多元化、随时随地、经济系统和文明。

因此，元宇宙就像一个“大箩筐”，它包含许多应用的可能性。理想的元宇宙容许用户进行任何体验或活动，或者解决他们几乎所有的需求，所以在理想状态下，元宇宙可应用于任何事物。

表9：元宇宙的多种应用领域

领域	应用示例	示例详情
商业领域	Meta 的 Horizon Workrooms 微软的 Mesh	为员工提供虚拟办公环境，支持远程协作。减少居住在城市的必要性。
	虚拟会议与活动平台	如 Hopin 和 Remo 等平台支持大规模在线会议和活动。
教育领域	英伟达的 Omniverse Together Labs 的逼真化身技术	支持全球开发者实时合作，创作元宇宙内容。利用 AI 技术模拟历史人物。
	香港科技大学的 MetaHKUST 项目	提供沉浸式互动学习环境。
	虚拟实验室与实验环境	学生可以在安全的虚拟环境中进行化学、物理等实验。
房地产领域	拟真的虚拟房屋参观 NFT 虚拟房地产	允许购房者在线参观房产。如"Mars House"以 50 万美元售出的 NFT 房屋。
	虚拟建筑设计与规划	建筑师和设计师可以在虚拟环境中进行设计和规划。
音乐领域	在《我的世界》、《机器砖块》、《堡垒之夜》等游戏中举办的元宇宙演唱会	饶舌歌手崔维斯·史考特 2020 年的《堡垒之夜》演唱会收入 2000 万美元。
	虚拟音乐会平台	Wave 和 TheWaveVR 等平台支持艺术家举办虚拟音乐会。
游戏领域	Roblox、Fortnite、Minecraft、The Sandbox	Roblox：用户可以创建自己的游戏和体验，形成一个庞大的游戏社区。Fortnite：举办虚拟音乐会和电影放映。Minecraft：玩家可以在虚拟世界建造和探索。The Sandbox：玩家可以创建、拥有和货币化自己的游戏体验。
	虚拟体育竞技	如虚拟足球联赛或赛车锦标赛。
	电子竞技平台	支持玩家参与和观看电子竞技比赛。
	虚拟角色扮演	玩家可以扮演不同的角色，在虚拟世界中完成任务和探险。
整合领域	HTC VIVE 推出的 VIVERSE 元宇宙生态系统	包括 5G 产品、VR 设备以及 ENGAGE 和 VRChat 等合作伙伴。
	元宇宙社交平台	Decentraland 和 Sandbox 等平台支持用户创建和探索虚拟世界。
企业零售领域	虚拟版的实体设施提供逼真的网上购物体验	顾客可以参观虚拟购物商场，试穿真实尺寸的虚拟商品。
	虚拟品牌体验店	品牌可以在元宇宙中开设虚拟商店，提供沉浸式的购物体验。
旅游领域	虚拟旅游景点	旅行者可以在家中游览世界各地的著名景点。
	交互式历史文化体验	通过虚拟现实重现历史事件，提供互动体验。
医疗健康领域	虚拟治疗室	心理健康专家可以提供远程心理治疗。
	虚拟康复训练	患者可以在家中进行物理治疗和康复训练。

资料来源：国信证券经济研究所整理

可以肯定的是，由于元宇宙的概念没有边界感，因此它的扩展性、延伸性很强，未来还将持续不断地扩展新场景，融合新技术，整合新功能，丰富新体验。

大型公司都希望在元宇宙上有所建树，Meta 就是个典型的例子。一方面他们可以投入很多在 VR/AR 硬件预算上，或者在软件算法上，或者像以前应用市场那样整合很多的资源，但有一个槛目前似乎还迈不过去：即如果他们承认区块链和 DAO 组织是本轮元宇宙最革命的要素之一，那么这些企业作为参与者的身份应该如何

定位？意思是：区块链如果像一新平台，这些大公司更像是这个新平台的一个 APP 而已。虽然他们不想放弃原有在互联网下的平台身份，但它们自身又缺乏颠覆性创新的要素。因此尽管这些大公司也在口口声声说自己要拥抱元宇宙，或者将元宇宙战略作为公司发展的最重要战略之一，但他们口中的元宇宙约等于一款新游戏，一个新办公软件，最多是一个新的应用商店，只不过多了一些 VR/AR 元素而已。

这就是互联网或者新技术最有魅力的地方——曾经的变革者通过 70、80 年代互联网浪潮逐步成了今天的垄断者，而新的变革又在挑战今天的垄断者，元宇宙这种颠覆式创新力量已经显现。

云计算的蓬勃发展

1、2016 年，全球主要互联网公司云计算业务均已上线

如果摩尔定律代表一种科技底层的驱动力量，那么伴随着 CPU 线程遇阻，云计算的发展成为了另一种自然的选择。

前期的报告提及过，亚马逊作为云计算最早的布局者，2006 年就推出了 EC2（Elastic Compute Cloud，弹性云）。随后谷歌在 2008 年推出了 App Engine，也开始云计算服务的试水。阿里巴巴则是中国最早开始布局云计算的公司，2009 年阿里云成立，并一直保持中国云计算市场份额第一的位置，微软 2010 年发布了 Azure 平台，苹果则在 2011 年推出了 iCloud。

国内巨头百度、中国电信、腾讯分别在 2012-2013 年发布了各自的云计算服务，IBM 则是通过收购在 2013 年推出了 IBM Cloud，甲骨文在 2016 年发布了基于 IaaS、PaaS 和 SaaS 的不同层次云计算服务。至此，在 2016 年，各大主流云计算平台均布局完毕。

表10：全球主要公司开展云计算的时间

公司名称	开始时间	备注
Amazon Web Services	2006 年	AWS 是亚马逊公司推出的云计算服务平台，被认为是最早的公有云服务之一。
Google Cloud Platform	2008 年	谷歌云平台最初面向开发者提供有限的服务，如 App Engine，2011 年 11 月开始全面向公众开放。
阿里巴巴	2009 年	阿里云成立于 2009 年，2011 年开始大规模对外提供云计算服务。
Microsoft Azure	2010 年	微软 Azure 最初名为 Windows Azure，于 2010 年 2 月开始提供服务。
华为	2011 年	华为 2011 年建立企业云业务部，全面试点公有云运营。
苹果 iCloud	2011 年	iCloud 于 2011 年 10 月发布，在推出后一周内就拥有了 2000 万用户。
百度	2012 年	百度云（现称为百度网盘）最初于 2012 年 3 月 23 日正式推出。
中国电信	2012 年	2012 年 3 月成立中国电信云计算公司，这是国内首家运营商级别的云计算公司。
腾讯	2013 年	2013 年 9 月开始对外提供服务。
IBM Cloud	2013 年	2013 年 6 月，IBM 收购了公共云平台 SoftLayer，作为其 IaaS 产品的基础。
Oracle Cloud	2016 年	2016 年 10 月，Oracle Cloud 是甲骨文公司推出的云计算服务，提供 IaaS、PaaS 和 SaaS。

资料来源：各公司网站，国信证券经济研究所整理

云计算的范围太过广泛，难以做出一般性定义。2011 年 NIST（美国国家标准与技术研究所）对云计算确定了“五个基本特征”：按需自助服务、广泛的网络访问、资源池化、快速弹性、可测量的服务。

此外，在讨论云计算市场规模时，在早期往往将互联网公司的广告收入也纳入进来（由于客户访问了网站的广告，而广告是在云端的服务），但后来人们逐渐意识到这并未有效区分商业模式的变化，因此此后的市场规模不再包含广告收入，即定义也随着时间的发展而有所变化。

目前，Gartner 将云计算重新分成四大类业务，业务流程即服务(BPaaS)、软件即

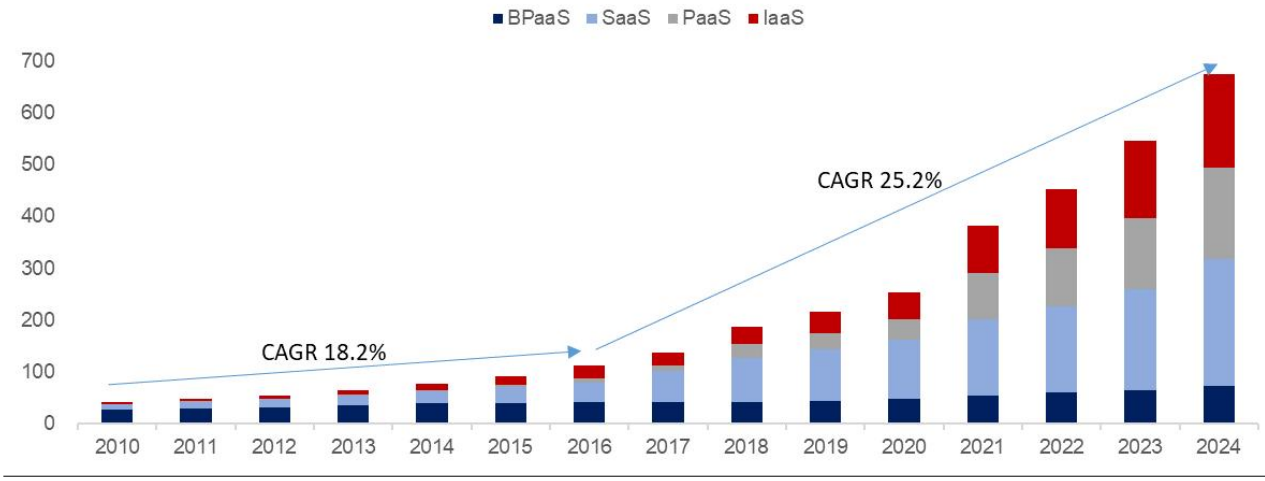
服务(SaaS)、平台即服务(PaaS)、基础设施即服务(IaaS)。

2010 年云计算的市场规模达到了 410 亿美元，2016 年市场规模超过 1120 亿美元，复合增速为 18.2%。进入 2016 年，随着多家互联网企业的布局并发力，云计算迎来了快速增长期，到了 2024 年，市场规模达到了 6760 亿美元，复合增速达到了 25.2%。

其中，2016-2024 年，BPaaS 复合增速 7.5%，SaaS 复合增速 25.9%，PaaS 复合增速最高，达到了 49.2%，IaaS 复合增速 28%位列第二。

从绝对规模而言，SaaS 体量最大，2024 年为 2440 亿美元；PaaS 和 IaaS 规模相似，分别为 1760 亿与 1820 亿美元。

图18：云计算市场规模，十亿美元



资料来源：wind，国信证券经济研究所整理

尽管从业务形态上，云计算区分了软件（SaaS）、基础设施（IaaS）、平台（PaaS）即服务，但是从企业发展的视角，无论是亚马逊、微软、谷歌还是阿里巴巴，它们都为客户提供一揽子解决方案，囊括了以上三者。

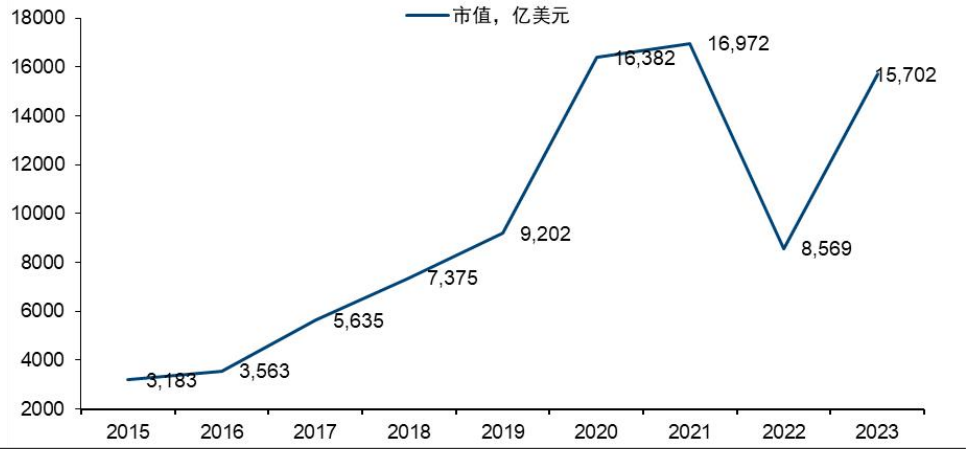
2、从云计算看投资价值

亚马逊作为云计算的引领者，AWS 业务在 2015 年 Q1 开始盈利，2016 年增速高达 54.4%，这比公司的主营收入增速高很多（公司主营收入增速：2015 年 20.2%，2016 年 27.1%，2017 年 30.8%）。或者可以说，是因为云计算高速增长，带动了公司收入的快速增长，云计算无疑成为公司的重要增长点。从 2016 年开始，市场开始为亚马逊的云计算估值，由于市场预估云计算业务成熟后净利润率可达 30%，当时给到云计算业务 10-15 倍的市销率（P/S），也就是 1220-1830 亿美元的估值，这几乎达到了公司市值的 1/3-1/2（2016 年底，亚马逊市值 3563 亿美元）——这瞬间抬高了其市值的天花板。市场突然意识到，这个从前没有怎么留意，仅占公司收入 9%的板块，居然可以贡献公司三分之一甚至一半的市值，更为重要的是，它的增速还特别高！即凭借着高增长的云计算，将拉动公司市值高速增长 3-5 年甚至更长的时间！

市场判断事后证明是对的。2016-2020 年，亚马逊的市值比 2015 年底翻了 5.1 倍，年化回报高达 39%。在 2022 年，当时亚马逊的市值已经跌落至 1 万亿美元左右，但有投行发布报告称亚马逊的云计算业务未来估值可达 3 万亿美元。（https://www.techweb.com.cn/ucweb/news/id/2896322_1）

因此，在资本市场上，2016 年可称得上云计算投资的元年。这个概念不再是缥缈的技术名词，而是实实在在支撑股价上涨的科技大方向。

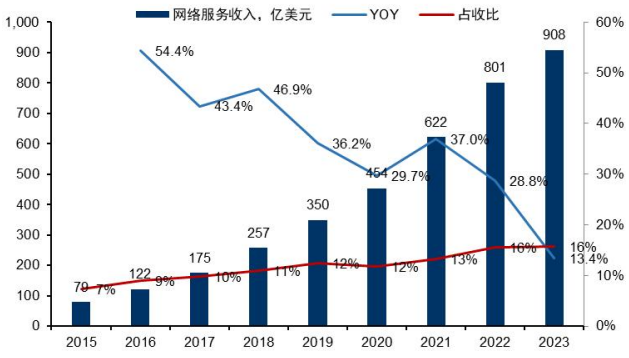
图19：亚马逊的市值



资料来源：wind，国信证券经济研究所整理

类似的情景也发生在微软、谷歌、阿里等公司上。微软公司除了 Azure，自身也在积极推进 Office 365 的渗透，加之企业端的 Dynamics 365，因此微软的云计算占收比更高，下图可见，云收入增速在 2017-2022 年呈现显著的加速的局面。2016-2021 年，微软公司也迎来了市值的大爆发，6 年时间翻了 4.7 倍，在 2021 年底达到了 2.5 万亿美元。

图20：亚马逊 AWS 收入、增速及占收比



资料来源：亚马逊，国信证券经济研究所整理

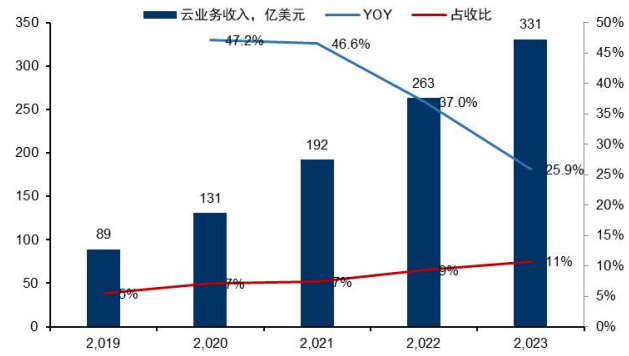
图21：微软智能云收入、增速及占收比

资料来源：微软，国信证券经济研究所整理

谷歌的云业务体量大约为亚马逊同期的 1/3，且谷歌的云业务收入达到 100 亿美元的时间已经到了 2020 年。这使得谷歌的云计算形成规模体量的时间要晚一些，使得其对谷歌市值的推动作用也要小一些。

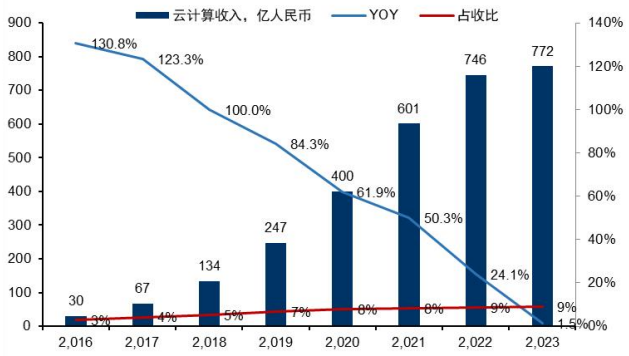
相对于三家美国公司，阿里巴巴的云计算收入在主营收入中的占比是最低的，但增速却是最快的。例如，2016 年，阿里云收入增速 130%，占收比 3%；到了 2020 年，阿里云的收入达到了 400 亿元人民币，增速 61.9%，占收比 8%。按照 15 倍左右的市销率，阿里云在 2020 年的估值大约在 800 亿美元，约占到总市值的 12%。但由于新冠疫情产生后，电商是居家办公的主要受益者，加之美联储降息，阿里巴巴在 2020 年依然获得了市场的追捧，市值较 2015 年底翻了 2.2 倍，达到了 6507 亿美元。

图22：谷歌云服务收入、增速及占收比



资料来源：谷歌，国信证券经济研究所整理

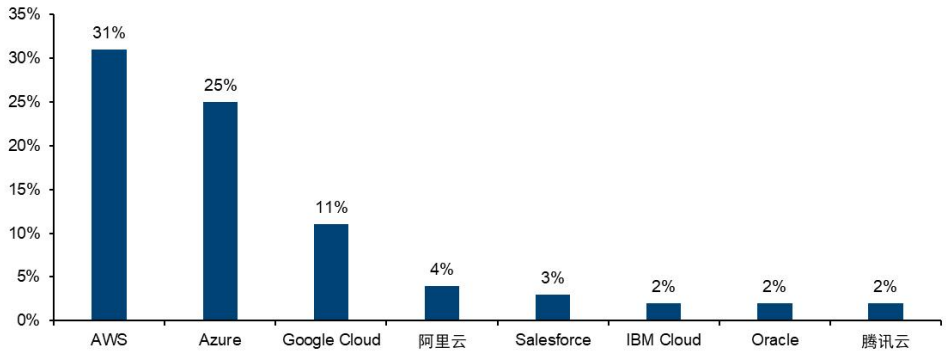
图23：阿里巴巴云计算收入、增速及占收比



资料来源：阿里巴巴，国信证券经济研究所整理

综上，无论是亚马逊、微软、谷歌还是阿里巴巴，在大力发展云计算业务时期，都极大扩展了其业务的天花板，并在主营业务中增加了新的增长极，都直接对市值形成了非常明显的正向作用。

图24：2024 年第一季度，全球云计算市场份额



资料来源：statista.com，国信证券经济研究所整理

但需要指出的是，时间到了 2023 年之后，各家公司的云计算业务增长都明显放缓，其中亚马逊、微软增速低于 20%，谷歌 25%，阿里巴巴由于一些原因增速很低。因此云计算的最大红利期已经过去，这些企业亟待寻找云计算之外的新业务增长点。

新能源汽车

1、为什么新能源汽车是“行走的电脑”？

依然，如果我们将摩尔定律看做是一种永不停歇的、科技发展的底层驱动力，那么新能源汽车的第一特征可能不再是汽车，而是一台“大号的、行走的计算机”。

由于这台“计算机”功耗比较大，所以它需要更强大的供电系统，而为了使它拥有更强大的供电系统，就必须将传统的汽油动力改造成新能源动力。

目前我们能够看到的汽车在未来最重要的特征是无人驾驶。为什么传统汽油车没有办法实现无人驾驶？

1、电力供应限制：传统汽油车的电力系统主要是为启动发动机和一些基本的电子设备供电设计的，通常使用的是 12 伏特的铅酸电池，这种电池的容量较小，不足以支持无人驾驶系统所需的大量电力；而无人驾驶系统需要大量的电力来驱动高

性能的计算单元、传感器（如激光雷达、毫米波雷达、摄像头等）以及各种电子控制单元，这些设备的总耗电量远超过普通汽油车的电力系统所能提供的；

2、电源稳定性：汽油车的发电机输出功率和电压并不总是稳定的，这可能导致无人驾驶系统的可靠性降低。例如，当发动机转速变化时，发电机的输出也会随之变化，这对于需要高度稳定电源的无人驾驶系统来说是一个潜在的问题；电动汽车的电力系统更加稳定，因为电池组可以提供更平滑的电流输出，这对于敏感的电子设备来说非常重要；

3、内部数据架构：由于无人驾驶系统需要与车载电子设备进行高效互联，而传统汽油车的电气架构可能并不支持这种高集成度的数据传输和处理。即便后装相关设备，也可能因设计限制导致电路过载、线缆杂乱等问题，影响性能和安全性；

4、软件与 OTA：传统汽车的软件几乎不需要更新，即便更新，也是以年度计的，或者是在汽车出了特定问题后维修或者厂家召回时才进行软件升级；而新能源电动汽车的运行机制与电脑、PAD、手机没有区别，软件通过 OTA 的方式升级，由于它可以实时通过大数据去积累参数，优化体验，OTA 需要周度甚至日度更新，这需要车辆具备 OTA 实时连接的能力；

5、车辆结构和空间限制：无人驾驶系统需要安装大量的传感器和计算设备，这些设备需要特定的空间布局。传统汽油车的设计往往没有预留足够的空间来容纳这些额外的设备。汽油车的机械结构（如发动机、传动系统等）占据了较大的空间，留给新技术的空间相对有限。

6、未来可能会出现硬件升级方案。如果把传统汽车比作“白电”，新能源电动车则是摩尔定律的产物，是“黑电”。特征是，即便用户买到最先进的型号，也会在使用不长时间之后变得落后（想象一下用了三年的智能手机）。因此未来的新能源汽车的设计可能是类模块式的，车企可以通过收服务费或者升级费的方式来升级部分无人驾驶相关的配件。部分厂家已经推出了类似的方案，但这依然需要观察。

2、从 Model-S 开始新能源车进入快车道

广义的新能源汽车的发展，甚至拥有了百年的历史。但是我们不打算回顾这段历史，因为尽管它们是以电力驱动的汽车，但给人的印象是：速度慢，加速时间长，行驶里程短的印象。

90 年代之后，通用汽车公司的 EV1 与特斯拉公司的第一代 Roadster，是现代新能源汽车领域的两次比较重要的尝试。

EV1 是当代第一辆由主要汽车制造商批量生产和专门设计的电动汽车，1996 年推出并于 1999 年停产。与当时众多的电动汽车不同，EV1 是一款专门打造的电动汽车，而非其他汽车的改装版。这一因素导致其研发费用高达 3.5 亿美元，生产成本也高昂。

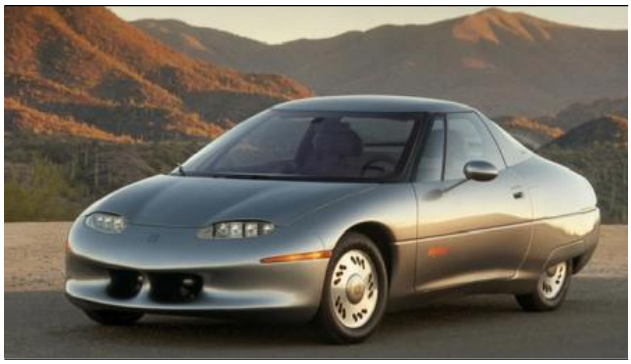
1996 年上市的第一代 EV1 由铅酸电池供电，续航里程为 70 至 100 英里，一共生产了 660 辆，颜色包括深绿色、红色和银色，这些汽车通过租赁方式提供，不能购买（虽然有标价 34,000 美元）。EV1 的租赁价格从每月 399 美元到 549 美元不等，最初的承租人包括名人、企业高管和政客等知名人士。在上市的第一年，通用汽车仅租赁了 288 辆汽车。

1998 年，通用汽车发布了第二代 EV1。改进包括降低生产成本、运行更安静、大幅减轻重量以及推出镍氢电池(NiMH)。使用铅酸电池组的汽车行驶里程为 80 至

100 英里(130-160 公里),而镍氢电池组汽车每次充电可行驶 120 至 140 英里(190-230 公里)。第二代 EV1 租赁计划扩展到其他几个美国城市，每月支付的费用从 349 美元到 574 美元不等。通用汽车共生产了 457 辆第二代 EV1。

尽管客户反响良好，但通用汽车认为电动汽车市场无利可图。通用汽车根据 33,995 美元的初始车辆价格确定了 EV1 的租赁费用。但业内人士估计每辆 EV1 的成本约为 8 万到 10 万美元。美国《史密森尼杂志》将 EV1 描述为“从技术上来说不算失败”，《澳大利亚金融评论报》则认为，虽然“EV1 很成功，但注定会失败”。这些观点是因为 EV1 在经济上不可行，而通用汽车因停止生产 EV1 而受到认可，《汽车新闻》声称这一决定帮助通用汽车避免了数十年的亏损。

图25：通用 EV1（1996 年）



资料来源：百家号，国信证券经济研究所整理

图26：特斯拉 Roadster（2008 年）

资料来源：车家号，国信证券经济研究所整理

EV1 的消亡启发了特斯拉。马丁·埃伯哈德和马克·塔彭宁于 2003 年 7 月成立了特斯拉汽车公司。半年后，埃隆·马斯克入股并担任董事长一职。公司的战略是从针对早期采用者的高档跑车开始，然后转向更主流的车型，包括轿车和价格实惠的紧凑型车。

Roadster 是一款基于莲花跑车（Lotus Elise）底盘设计的电动跑车。在 2008 年至 2012 年间生产,它是第一款在高速公路上合法行驶的使用锂离子电池的量产全电动汽车，也是第一款每次充电行驶超过 244 英里（393 公里）的量产全电动汽车。根据型号不同，该车可在 3.7-3.9 秒内加速百公里/小时，最高时速为 125 英里/小时(201 公里/小时)。它的电池到车轮的能耗为 21.7 千瓦时/100 英里(135 瓦时/公里），平均效率为 88%。

由于是初创公司，Roadster 在研发过程中遭遇到很多麻烦，如：1、两家变速箱供应商无法制造出所需的配件，公司在量产后宣布变速箱可靠性存在问题；2、特斯拉 AC Propulsion 的动力传动系统交给 Lotus 公司进行代工生产，但大家都忽略了在原有的莲花跑车底盘上装载沉重的动力电池不可能不改变底盘设计；3、项目的研发成本一提再提，最初的研发预算大约为 2500 万美元，而最终估计可能超过 1.05 亿美元，为特斯拉成功完成了多轮融资，且不得不提高售价；4、两次产品安全召回：分别是后内轮毂法兰螺栓问题与冗余备用系统的 12V 低压辅助电缆问题。

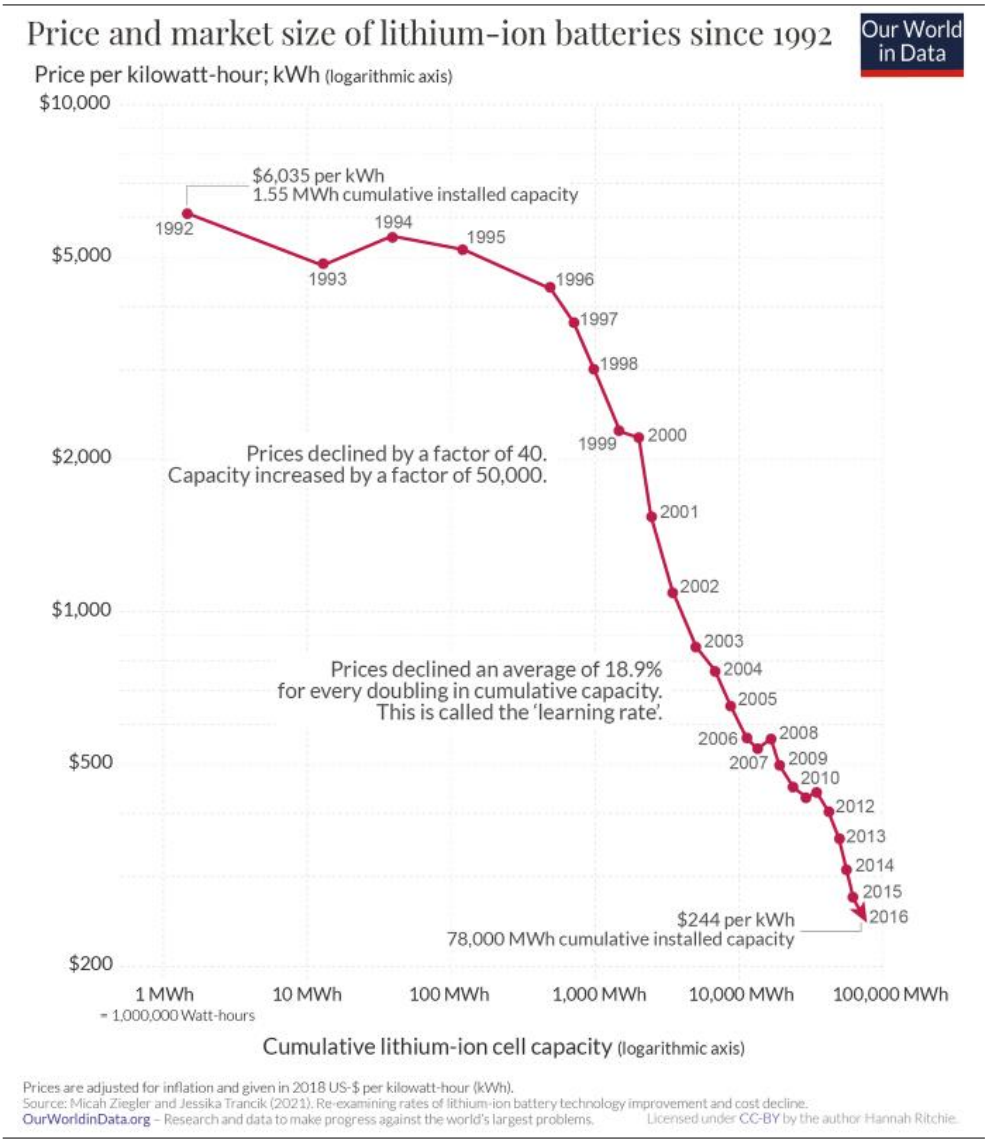
不过，通过多轮融资和马斯克个人的资金支持，特斯拉最终成功完成了 Roadster 的研发和生产，并于 2008 年 2 月至 2012 年 12 月期间，在全球交付了约 2,450 辆 Roadster。Roadster 虽然产量不大，但却验证了商业模式的可行。从这个角度说，它是特斯拉成功的一款产品。2010 年 5 月，特斯拉以 4200 万美元的价格从丰田手中收购了位于加州弗里蒙特的 NUMMI 工厂。2010 年 6 月，公司通过 IPO 在纳斯达克上市,这是自 1956 年福特汽车公司 IPO 以来第一家在纳斯达克上市的美

国汽车公司，募集资金 2.26 亿美元。2012 年，特斯拉停止生产 Roadster，并将重心转向 Model-S 的生产。

有了 Roadster 的经验积累，自己的制造车间，以及募集资金，特斯拉于 2012 年推出全自主设计的 Model-S。Model S 于 2012 年 6 月在加利福尼亚投产。Model S 的车身和底盘主要由铝制成，而其感应电动机则由钢和铜制成。该车在生产过程中进行了多次更新，2015 年和 2016 年，Model S 是世界上最畅销的插电式电动汽车，直到被特斯拉 Model 3 超越。这款车赢得了无数赞誉，包括被《时代》杂志评为“2012 年最佳 25 项发明”之一，并获得 CNET 颁发的“2012 年度科技汽车”奖。

2015 年 6 月，特斯拉宣布 Model S 行驶里程超过 10 亿英里，这是第一款达到这一总里程的全电动汽车。当年，Model S 的全球销量突破 10 万辆，2016 年 11 月突破 15 万辆。2017 年第四季度，Model S 销量突破 20 万辆大关。正是因为月销量在万级水平，产业链开始快速成熟，我们可以从电池成本的降幅上窥豹一斑。

图27：1992-2016 年锂电池成本变化，美元/千瓦时

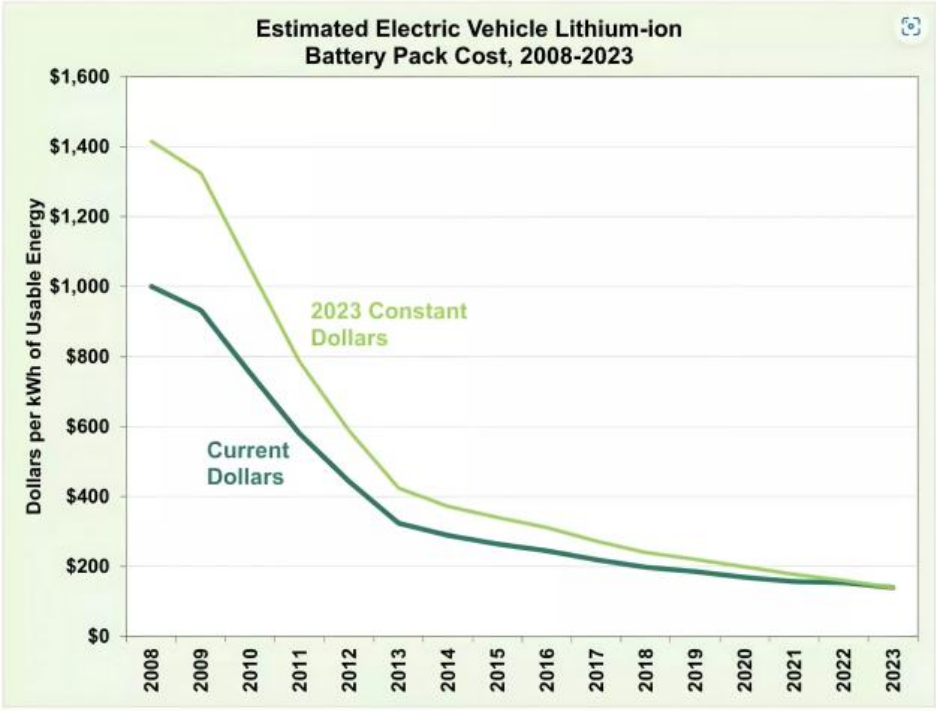


资料来源：OurWorldinData，国信证券经济研究所整理

图中可以看出，在通用汽车 EV1 时代（90 年代末），锂电池每千瓦时成本还在 3000-4000 美元，如果当时选用锂电池方案，哪怕用 Model S 最低配电池方案，40kwh，光电池成本也要 12 万-15 万美元；如果对标 Model S 中配 80kwh 方案，则电池成本也要 25 万-30 万美元！所以 EV1 当时只能选择镍氢电池。而从特斯拉诞生，到 Model S 的推出，锂电池成本从 2002 年的 1000 美元/kwh 快速降低到 2016 年的 244 美元/kwh，14 年时间里降幅累计约 75%。

如果查看 2016 年之后的成本下降速度，并未有此前那么剧烈，到 2023 年，每千瓦时成本已经降至 200 美元以下。BloombergNEF 的 2023 年度电池价格调查显示，锂电池组成本在 2023 年下降了 14%，达到每千瓦时 139 美元的历史新低，并预计到 2030 年降至 80 美元/千瓦时。如果按照 2016 年 244 美元/kwh 以及 2030 年 80 美元/kwh 计算，在后续这 14 年的时间里，降幅约 67%。假如到那时，我们即便配置 200kwh 的电池，其电池成本为 1.6 万美元，这个配置足以消除里程焦虑（如问界 M9，98kwh，纯电续航 630 公里，那么 200kwh 续航可达到 1200 公里以上）。

图28：2008-2023 年锂电池成本变化，美元/千瓦时



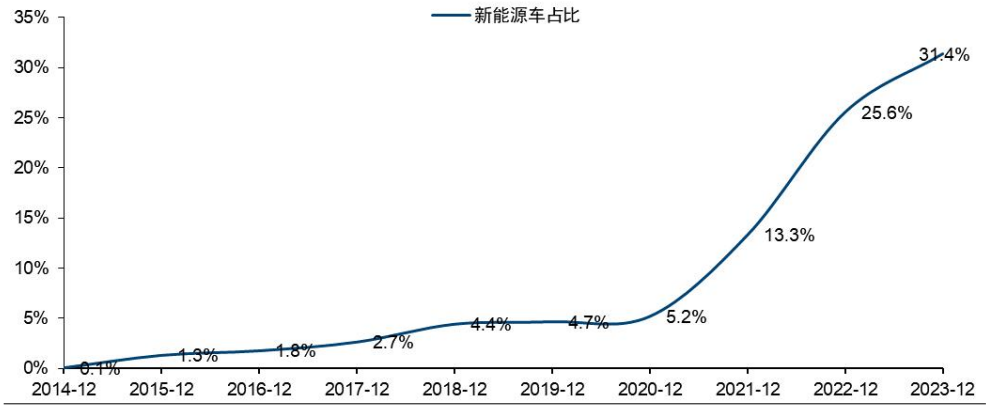
资料来源：199it.com，国信证券经济研究所整理

从以上案例可以发现，特斯拉进入到这个产业的时机很好。

通用的 EV1 进入时间显得过早，得到的反馈类似于寓言《小马过河》中松鼠的经验：“水太深，足能把人淹死！”而如果在 2020 年电池成本降至 200 美元/kwh 再选择进入市场，虽然得到的反馈类似于老牛的经验：“水很浅，一蹚就过去了！”但问题是，面临着诸多已经积累了几年、十几年的车企，自身的优势又如何积累呢？

无论是中国政府还是企业，均非常重视新能源车带来的巨大机会，新能源车产业链在中国获得了长足的发展。按照年度计算，2023 年新能源车占汽车的销售比例已经上升至 31.4%，如果按照月度数据计算，截至 2024 年 6 月，这个比重已经高达 41.1%。

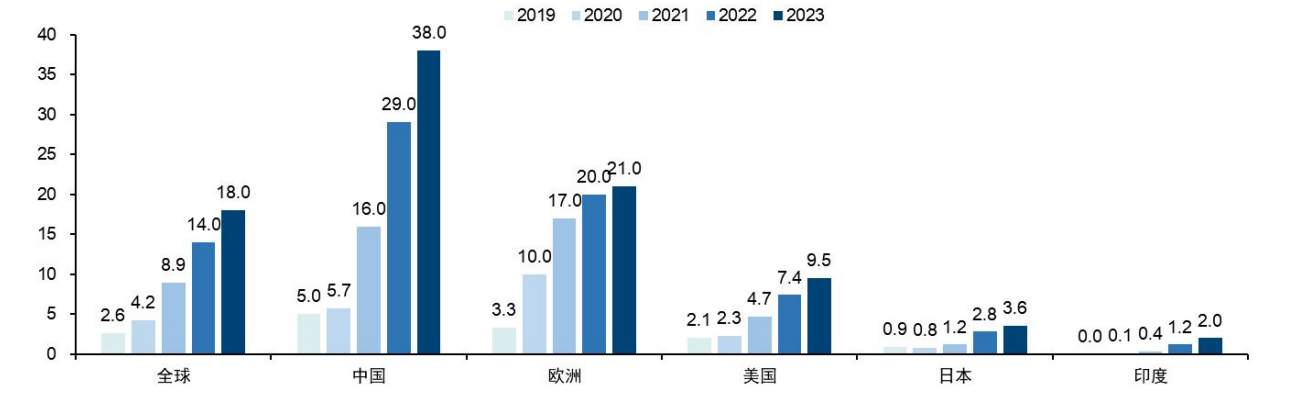
图29：中国新能源车销量占比



资料来源：中国汽车工业年鉴，国信证券经济研究所整理

如果放眼全球，中国新能源车的普及率远超过其他国家。下图可见，IEA 统计的全球新能源车销量占比在 2023 年达到了 18%，而中国超出全球平均水平 20 个百分点。欧洲 21%，美国为 9.5%，日本和印度仅为 3.6%和 2.0%。

图30：全球新能源车销量占比

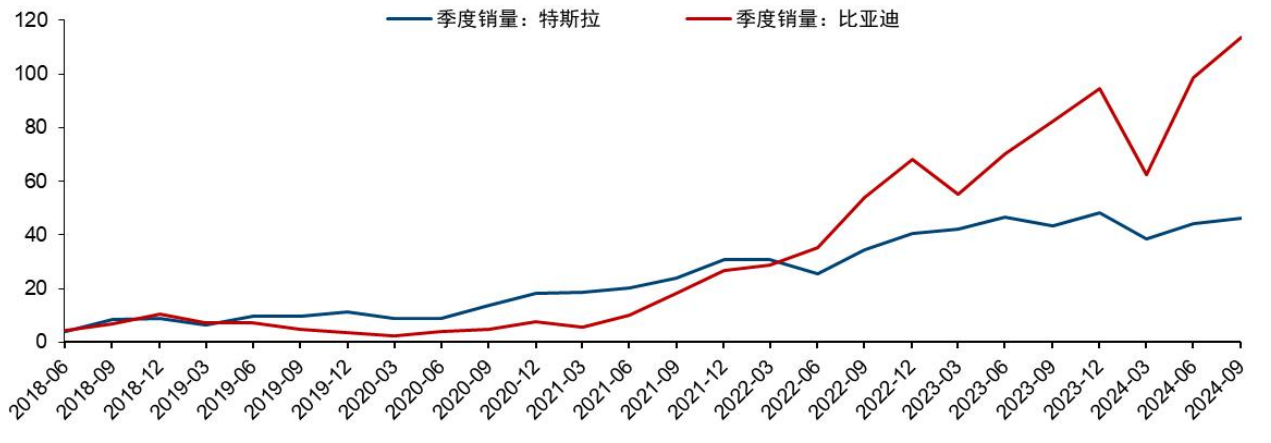


资料来源：IEA，国信证券经济研究所整理

在中国，比亚迪是最早进入新能源车领域的车企，目前市场份额第一名。除了比亚迪，还有造车新势力、吉利、上汽、塞力斯等多个车企也获得了快速的增长。

行业发展的缩影也投射到了龙头企业上。在 2022 年 Q1，比亚迪新能源车销量超过特斯拉，到了 2024 年 Q3，比亚迪季度销量已经达到 113 万辆，是同期特斯拉的 2.4 倍以上（46 万辆）。

图31：新能源车季度销量：比亚迪 VS 特斯拉（万辆）



资料来源：比亚迪、特斯拉，国信证券经济研究所整理

新能源车的普及大大推动了车载芯片的需求。以英伟达为例，其从 2015 年开始发布车载芯片，从最开始的算力 1T，到目前的 400T（应用在具体车型上大约在 250T），车载芯片的算力有了显著的提升。根据英伟达 2022 年 9 月发布的下一代智驾 Drive Thor 来看，届时算力将达到 2000 TOPS。华为曾预计，到 2030 年，平均每台新能源车的整车算力将达到 5000 TOPS。由于在 2023 年问界、2024 年百度出租车已经发布了无人驾驶方案并且已经在规模应用场景中取得了良好的反馈，在 2023-2024 年，车载芯片如果按照配置 2-4 块 Drive AGX Orin 芯片计算，算力大约在 500-1000T，故而随着算力的提升与数据训练，相信未来无人驾驶的体验将更加美好。

表11：英伟达车载芯片家族

芯片/模组名称	发布时间	交付时间	架构	制程	算力	TDP
Drive	2015 年 1 月	NA	Maxwell	28nm	NA	20W
Drive PX 2（Tesla）	2016 年 10 月	NA	Pascal	16nm	4 TFLOPS（FP32）	40W
Drive PX 2（Tesla 2.5）	2017 年 8 月	NA	Pascal	16nm	4 TFLOPS（FP32）	60W
Drive PX Xavier	2017 年 1 月	NA	Volta	12nm	20 TOPS（INT8）	30W
Drive PX Pegasus	2017 年 10 月	NA	Volta	12nm	320 TOPS（INT8）	500W
Drive AGX Orin	2019 年 12 月	2022	Ampere	8nm	400 TOPS（INT8）	130W
Drive AGX Pegasus OA	2019 年 12 月	2022	Ampere	8nm	2000 TOPS（INT8）	750W
Drive Atlan（已取消）	2021 年 4 月	NA	Ada Lovelace	4nm	1000 TOPS（FP8）	NA
Drive Thor	2022 年 9 月	2025	Blackwell	4nm	2000 TOPS（FP8）	NA

资料来源：维基百科，国信证券经济研究所整理（算力应用到具体车型时会与发布时有差异）

大模型的出现：AI 翻开了崭新的一页

Transformer 架构的出现

从 1966 年的 ELIZA 开始到 2022 年的 56 年时间里，计算机科学家想尽各种办法设计了多个聊天机器人，也发布了各种版本的翻译软件，但它们的聊天体验总是有些差强人意。直到 2022 年 11 月 30 日，OPENAI 公司发布了基于大语言模型的 ChatGPT 聊天机器人：它的聊天体验过于流畅与丝滑，马上成为全球科技界的现象级产品。

到 2023 年 1 月，它已成为当时历史上增长最快的消费软件应用程序，吸引了超过 1 亿用户。ChatGPT 的发布刺激了竞争产品的发布，包括 Gemini（谷歌）、Claude（Anthropic）、Llama（脸书）、文心一言（百度）和 Grok（xAI）等，并引发了科技企业在 AI 大模型方向上的军备竞赛。

表12：历史上典型的聊天机器人

名称	发布年份	开发者	简介	架构
ELIZA	1966	Joseph Weizenbaum	最早的自然语言处理程序之一，通过简单的模式匹配进行对话。	模式匹配，基于规则的对话生成
PARRY	1972	Kenneth Colby	模拟偏执狂患者的聊天机器人，使用模式匹配进行对话。	模式匹配，模拟特定人格特质
ALICE	1995	Michael Mauldin	基于 AIML 的聊天机器人，能够理解简单的自然语言指令并给出回答。	AIML（人工智能标记语言），基于规则的框架
Mitsuku（Kuki）	2005	Steve Worswick	5 次获得勒布纳奖的聊天机器人，能够进行复杂的对话。	AIML 扩展，使用模式匹配和数据库
Cleverbot	2008	Rollo Carpenter	利用机器学习技术不断改进对话能力。	早期的机器学习，使用统计方法分析用户输入
Watson	2010	IBM	强大的认知计算系统，在电视节目中击败人类冠军。	混合架构，包括规则基础系统和自然语言处理技术
Xiaoice	2014	微软	针对亚洲市场设计的社交聊天机器人。	深度学习，使用神经网络模型进行对话生成
Replika	2017	Replika Labs	个性化的人工智能伴侣，可根据用户的个性进行定制。	深度学习，使用神经网络模型进行个性化对话生成
BlenderBot	2020	Facebook AI Research	开源的对话生成模型，能够生成更自然流畅的对话。	预训练模型，使用 Transformer 架构
ChatGPT	2022	OpenAI	大型语言模型，能够生成高质量的文本和对话。	Transformer 架构的大规模预训练模型

资料来源：百度百科，国信证券经济研究所整理

2014 年，一个模拟 13 岁乌克兰男孩的计算机程序“Eugene Goostman”在英国通过了图灵测试。但批评者认为，假装自己是一名 13 岁的男孩是一种诡计，这会限制提问者的谈话，从而提高比分。但对于 ChatGPT 4.0 版本，2023 年 7 月 Celeste Biever 在《自然》杂志写道“ChatGPT 打破了图灵测试”；2024 年 3 月斯坦福大学的研究人员也称 ChatGPT-4“通过了严格的图灵测试，与普通人类行为的不同之处主要在于更加合作”。换句话说，如果标准相对严苛，ChatGPT 是迄今为止第一个被广泛认同的通过图灵测试的聊天机器人。

Transformer 是基于 2017 年谷歌公司的论文《注意力就是你所需要的一切》（Attention Is All You Need）中提出的多头注意力机制。文本被转换成称为“token”的数值，每个 token 通过查找词嵌入表被转换为向量。在每一层，每个 token 都会在上下文窗口范围内与其他 token 进行语境化，语境化的过程就是向量权重的计算过程，通过计算查询向量（Query）、键向量（Key）和值向量（Value）之间的相似度来实现，从而放大关键 token 的信号，减弱不太重要的 token 的信号。

Transformer 最初是作为对以前机器翻译架构的改进而开发的，但由于它有几个以前算法不具备的优势，它们被用于大规模自然语言处理、计算机视觉（视觉变换器）、强化学习、音频、多模态处理、机器人技术甚至下棋。它还导致了预训

练系统的发展。Transformer 架构的优势主要包括：

1、并行处理能力：可以通过自注意力机制同时关注序列中的所有位置而非 RNN 网络那样按照顺序处理；

2、长距离依赖关系的捕捉：通过自注意力机制能够有效地捕捉输入序列中的长距离依赖关系而非仅仅局部上下文；

3、灵活性和可扩展性，通过增加模型的层数或者改变每层的宽度可以轻松调整模型的容量，也可以很好地扩展到更大的数据集上，这有助于构建更大更复杂的模型；

4、通用性：不仅适用于机器翻译这样的特定任务，还可以应用于广泛的 NLP 任务，包括文本分类、情感分析、问答系统等；此外，Transformer 架构不仅仅限于 NLP 领域，它还被成功地应用于计算机视觉、音频处理等领域；

5、迁移学习和预训练：模型特别适合于迁移学习和预训练技术。通过在一个大型语料库上进行无监督预训练，然后在特定任务上进行微调，可以显著提高模型的性能。

因此，Transformer 架构凭借其高效、灵活、通用等特点，在 NLP 领域占据了主导地位，并且随着研究的深入和技术的进步，它的应用范围仍在不断扩展。《Attention Is All You Need》论文可谓是影响力巨大，截至 2023 年，所有 8 位作者都离开了谷歌，并创办了自己的 AI 初创企业，其中Łukasz Kaiser 加入了 OpenAI）。

GPT 与 BERT：让 Transformer 架构一举成名

OPENAI 公司成立于 2015 年，初期它的 GPT 并非一举成名，由于谷歌的近水楼台（2017 年发布的 Transformer 论文），BERT 模型于 2018 年 10 月由谷歌的 Jacob Devlin 等人在论文《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》中提出的。

BERT 模型是基于 Transformer 模型的一种变体，并且进行了多方面的改进和创新，包括双向编码、掩码语言模型、下一句预测任务、微调等等，最重要的是，Transformer 模型通常需要大量的标记数据来训练，而 BERT 则通过无监督的方式预先训练了一个大型语言模型。BERT 最初以英语实现，有两种模型大小：BERT BASE（1.1 亿个参数）和 BERT LARGE（3.4 亿个参数）。两者均在 Toronto BookCorpus（8 亿个单词）和英文维基百科（25 亿个单词）上进行训练。2020 年 3 月，发布了 24 个较小的模型，最小的是 BERT TINY，只有 400 万个参数。

BERT 在 4 个 Cloud TPU（总共 16 个 TPU 芯片）上训练 BERT BASE 耗时 4 天，预计成本为 500 美元。在 16 个 Cloud TPU（总共 64 个 TPU 芯片）上训练 BERT LARGE 耗时 4 天。因此，至少在这一时期，定位于翻译软件的 BERT 模型，还没有将后续的“大模型”以及巨额算力投入的特征体现出来。但即便如此，BERT 因与之前最先进的模型相比有了显著的改进而引人注目。

此时，创业中的 OPENAI 公司，如果想做出与 BERT 模型不一样的能力，只有一条路，就是增加训练层数，增加向量维度，增加训练数据，增加参数数量，换句话说，就是做出更大的模型。但在那个时候，这条路的效果怎样，还不得而知。

GPT-1 在 2018 年 6 月面世，其参数数量为 1.17 亿，与 BERT 相比，并未显示出明显的优势；

GPT-2 于 2019 年 2 月面世，此时参数放大了十倍到 15 亿，层数到了 48，训练数据也增加了 10 倍。GPT-2 与其前身 GPT-1 以及后继者 GPT-3 和 GPT-4 一样，都具有生成式预训练 Transformer 架构，实现了深度神经网络，特别是 Transformer 模型，它使用注意力机制，而不是基于旧式递归和卷积的架构，该模型可以大大提高并行化程度，使得模型在短期内不断升级成为可能；

2020 年 5 月，GPT-3 则是将参数数量增加到了 1750 亿。从 3.0 版本开始，GPT 展现了一些数学运算、代码生成、文本改写、自动推理、跨语言翻译等特征，研究人员将这一现象称为“涌现”（Emergence），即模型展现出未曾明确编程的新能力的现象，这大约是 OPENAI 坚持走“大模型”之路的重要回报；2022 年 11 月，GPT-3.5 面向公众的发布，在技术界和普通用户中都引起了广泛的兴趣和讨论。面对 GPT 的功能，人们见识到了历史上从未体验到的顺畅交互与“涌现”能力；由于它带来的能力过于震撼，让竞争对手都觉得“大力出奇迹”可能是当代人工智能不得不迈过去的坎，错过对大模型能力的探索，可能会在 AI 领域满盘皆输。至此，Transformer 确定了其江湖地位，大企业也开启了 AI 的军备竞赛；

2023 年 4 月发布的 GPT-4，估计参数高达 1.76 万亿，其耗费大约 2.5 万个 A100 的 GPU 显卡，90-100 天的训练时间。此时 A100 显卡单块高达 1-2 万美元，2 万块显卡成本高达 4 亿美元，这还不包含服务器基础设施、网络设备、存储设备、场地、电力成本、人力成本。

表13：GPT 几个主要版本

版本	发布时间	参数数量	层数	备注
GPT-1	2018 年 6 月	1.17 亿	12	基于 Transformer 架构的单向语言模型，使用了约 4.5GB 的文本数据进行训练。
GPT-2	2019 年 2 月	15 亿	48	训练数据 40GB 文本，800 万个文档，来自 Reddit 上 4500 万个获得点赞的网页。
GPT-3	2020 年 5 月	1750 亿	96	训练数据 570GB 纯文本、3000 亿个 CommonCrawl、WebText、英文维基百科标记以及两本书语料库。
GPT-3.5	2022 年 11 月	1750 亿	96	GPT-3.5 是 GPT-3 的微调版本，通过 RLHF（人类反馈强化学习）进行了优化。
GPT-4	2023 年 3 月	17600 亿	120	使用混合专家模型（MoE）进行构建，在 2.5 万个 A100 GPU 上进行了 90-100 天的训练。

资料来源：维基百科，国信证券经济研究所整理

我们在前文阐述过，近 30 年每 GFLOPS 算力成本大约年化降幅是 42%-45%，或者 15 年前的成本大约是当下的 1000 倍。换句话说，即便在 2008 年前后，哪怕是乔布斯还活着，假定他读到了一篇来自未来的关于 Transformer 架构的论文，但受制于当时的算力成本，如果当时想获得 GPT3.5 大模型的所需要的算力，其算力成本高达 4 千亿美元，假定其他投资（服务器、网络、存储、场地等）合计成本等于或者高于 GPU，则总成本要到万亿美元。

这就是算法受制于算力基础能力的原因——不是前人在几十年 AI 的探索中没有想到像 Transformer 这样好的方法，而是即便有这样的方法，也负担不起这么高的算力成本！

文生图与文生视频：从文字走向多模态

Midjourney 是一款 2022 年 3 月面世的 AI 绘画工具，当比较 Midjourney 随着时间的推移对“哈利波特的超现实形象”这一提示的响应时，生成模型的巨大进步显而易见。

图32：哈利波特的超现实形象，在不同时间 Midjourney 的输出

资料来源：斯坦福大学，国信证券经济研究所整理

2024 年 2 月 15 日 OPENAI 公司发布了 Sora 模型。Sora 是一种扩散模型(diffusion)与 Transformer 模型的结合体，它首先从看起来像静态噪声的视频开始生成视频，然后通过多个步骤消除噪声来逐渐转换视频。Sora 能够一次性生成整个视频，也可以延长生成的视频以使其更长。从 OPENAI 发布的视频来看，它一般可以生成 1 分钟左右的视频，而这是以往 3-4 秒钟的视频有着巨大的进步。

使用者可通过输入提示词来生成对应的模型。如下：

视频提示描述：两艘海盗船在咖啡杯中航行并互相搏斗的逼真特写视频。

视频提示描述：一位时尚女性走在东京的街道上，街道上到处都是温暖的霓虹灯和动画城市标识。她穿着黑色皮夹克、红色长裙和黑色靴子，手拿黑色手提包。她戴着太阳镜，涂着红色口红。她自信而随意地走着。街道潮湿而反光，五颜六色的灯光营造出镜面效果。许多行人走来走去。

图33：OPENAI 的 SORA 模型（咖啡杯里的海盗船）

资料来源：OPENAI，国信证券经济研究所整理

图34：OPENAI 的 SORA 模型（东京街头的女子）

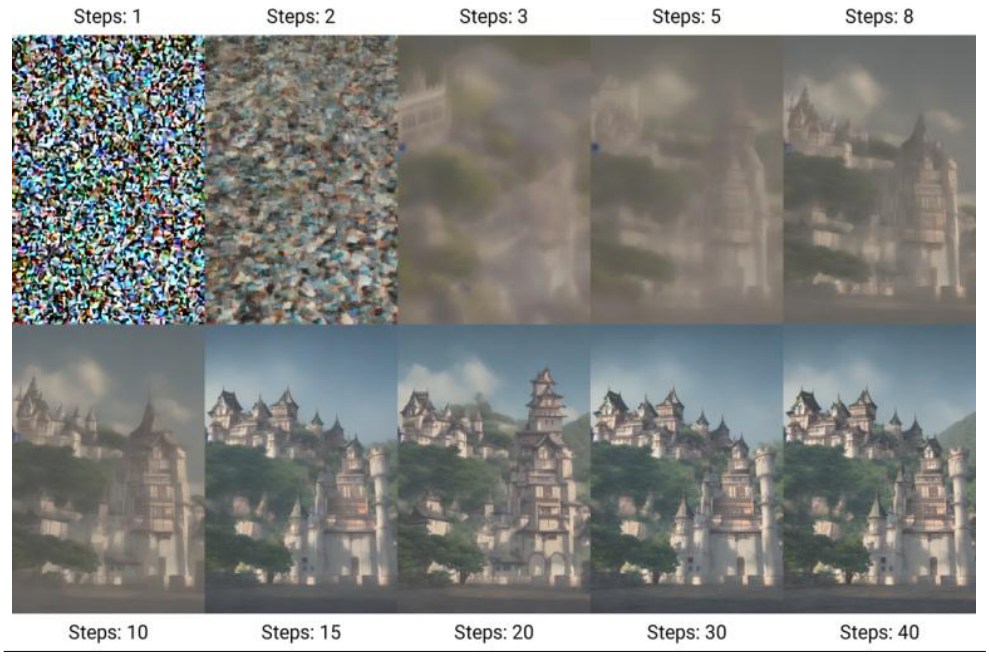
资料来源：OPENAI，国信证券经济研究所整理

例如，第二个视频中，无论视频中的女士行走在哪里，其背后的景物会跟随人物的变化而同步进行位置、光影的调整，这非常类似影视作品中的“运镜”。

2022 年，由慕尼黑大学 CompVis 小组和 Runway 公司的研究人员参与开发的 Stable Diffusion，一种基于扩散技术的深度学习文本转图像模型发布。2023 年 2 月，

Runway 发布了 Gen-1 和 Gen-2，这是第一个商业化、公开可用的基础视频到视频和文本到视频生成模型，可通过 Web 界面访问。

图35：扩散模型的去噪过程



资料来源：维基百科，国信证券经济研究所整理

扩散模型分成两个过程，前向扩散过程和反向生成过程。在前向扩散过程中，原始数据通过一系列步骤被逐渐加噪；反向生成过程则是从前向扩散过程的终点开始，逐步去除噪声并恢复到原始数据分布的过程。加噪与去噪都不是一次完成的，要通过多个步骤，去噪则可以引入神经网络来完成。因此扩散模型的训练目标是优化一个损失函数，使得模型能够学会如何从噪声逐步恢复到原始数据的分布。

例如，DALLE、DALLE 2 和 DALLE 3 是由 OpenAI 开发的文本到图像模型。与 GPT 相比，文生图并不需要很大的参数，例如市场对 DALLE 2 的反响不错，它的参数为 35 亿个，这就决定了其可以较早的投入商用，一些公司认为它可能代表未来数万亿美元产业的转折点。担忧的人则认为，将人工智能生成的图像与艺术家的作品并列是一种侮辱，破坏了艺术投入的时间和技能。此外是版权和文本到图像模型训练数据的问题，OpenAI 尚未公布用于训练 DALLE 2 的数据集的信息，即艺术家的作品在未经许可的情况下被用于训练。

图片提示描述：白色的基克拉迪房屋，点缀着蓝色的装饰和鲜艳的洋红色九重葛，坐落在宁静的希腊岛屿上。

图片提示描述：用纱线制作的海滩场景数字插图。沙滩用米色纱线描绘，海浪用蓝色和白色纱线拍打海岸。一轮纱线太阳落在地平线上，散发出温暖的光芒。纱线棕榈树轻轻摇曳，小小的纱线贝壳点缀在海岸线上。

图36：OPENAI 的 DALLE 3 模型（希腊小屋）

资料来源：OPENAI，国信证券经济研究所整理

图37：OPENAI 的 DALLE 3 模型（纱线质感的海滩）

资料来源：OPENAI，国信证券经济研究所整理

有了扩散模型，文生图变得简单，于是人们开始探索文生视频。它们的原理相似，只不过增加了时间维度数据。此时，模型学习的是把一组图片（而不是一张图片）去噪。这带来了显而易见的成本压力，因为训练过程的显存压力巨大，要投入巨大的硬件资源来支撑。此外，由于多了时间维度，在不同维度中精细调整画面的一致性也是挑战，如一个精细的人脸表情可能会因视频时间的延长而变得扭曲。在 SORA 问世之前，Stable Diffusion、Pika 等都是表现力不错的视频扩散模型。

VideoPoet 是 Google Research 于 2023 年为视频制作开发的大型语言模型。这与 OPENAI 公司的 SORA 模型类似，都继承了 Transformer 模型的特征，可以将文字、图片、声音、视频都转化为 token（文本、图片中的最小单位），再由 Transformer 模型来训练权重，这个过程非常类似于人类的理解过程，因此 VideoPoet 与 Sora 对视频或者多模态的理解力更强，同时在处理更长序列的数据也更有优势。但它们显然的问题是资源消耗，因为扩散模型可以将一幅图片通过几步就完成了降噪的过程，而 Transformer 模型需要将图片或者视频分割，而分割得越精细，则复杂度就会成倍的提升。

为了平衡更好的效果与更多的资源消耗之间的矛盾，目前 SORA 则是中和扩散模型和 Transformer 模型，将视频和图像表示为块（patches）的较小数据单元的集合，每个块都类似于 GPT 中的 token。通过统一数据的方式，SORA 可以在比以前更广泛的视觉数据上训练扩散 transformer 模型（diffusion transformers），涵盖不同的持续时间、分辨率和宽高比。

今天的 SORA，还存在着一些缺陷，下面是 OPENAI 公司发布的两个案例。

视频提示描述：一个人奔跑的步进场景，以 35 毫米拍摄的电影胶片。**图中的人向着跑步机反方向运动。**

视频提示描述：一位祖母，头发梳理整齐，灰白的头发，站在木制餐桌旁，五颜六色的生日蛋糕和无数根蜡烛后面，表情纯粹，快乐幸福，眼中闪烁着幸福的光芒。她倾身向前，轻轻一吹，蜡烛熄灭了，蛋糕上撒满了粉红色的糖霜和糖屑，蜡烛停止闪烁，祖母穿着一件饰有花卉图案的浅蓝色上衣，可以看到几个快乐的朋友和家人坐在桌边庆祝，但焦点模糊了。这个场景拍得很漂亮，很有电影感，展示了祖母和餐厅的 3/4 视图。温暖的色调和柔和的灯光增强了气氛。**图中的祖母吹完蜡烛之后，蜡烛仍在燃烧，但所有人包括祖母都表现得像是吹灭了一样。**

这些案例表明，在没有足够的训练数据时，SORA 也会对空间、时间的“理解”出现“错乱”。

图38：OPENAI 的 SORA 模型（反向的跑步机）

资料来源：OPENAI，国信证券经济研究所整理

图39：OPENAI 的 SORA 模型（吹不灭的生日蜡烛）

资料来源：OPENAI，国信证券经济研究所整理

由于扩散模型中的 CNN（卷积神经网络）在层数过多时就会出现梯度消失/梯度爆炸，这会使得模型“增益放缓”（Gain saturation）。而 Transformer 模型的特点是“大力出奇迹”（Scaling laws），数据量越大，参数越多，其训练效果则越好。一种极致的思考是：如果算力资源低廉得不需要计较，那么显然 Transformer 模型更优。因此，业界也普遍认为 Transformer 模型代表了未来。所以通过增加模型规模、使用更大的数据集以及更多的计算资源，来提高模型性能，而先不去考虑带来的增益水平是否值得，即先干了再说，可能已经成了行业普遍或者不得不认同的思路，而这种“互带节奏”的方式，正是当今文生图、文生视频、多模态行业的竞争现状。

2027 年 AGI 诞生？

OPENAI 宣称，Sora 为理解和模拟现实世界的模型（models that can understand and simulate the real world）奠定了基础，并相信这一能力将成为实现 AGI 的重要里程碑。

那么 AGI（Artificial general intelligence，通用人工智能）将在何时诞生？

按照维基百科的定义，AGI 是一种在广泛的认知任务中与人类智能能力范围相匹配的人工智能。创建 AGI 是人工智能研究以及 OpenAI 和 Meta 等公司的主要目标。

实现 AGI 的时间表仍然是研究人员和专家们争论的话题。截至 2024 年，一些人认为可能在几年或几十年内实现；另一些人则认为可能需要一个世纪或更长时间；少数人认为可能永远无法实现，而另一少数人则表示它已经存在。

随着 Transformer 模型商业应用的开展，打开的潘多拉魔盒再也无法合上，因为大公司已经嗅到只要按照“大力出奇迹”的路线不停地走下去，终将有一天 LLM 会达到，甚至超越人类的智商。

2024 年 6 月 OPENAI 公司的 CTO 米拉·穆拉提（Mira Murati）提到：“根据系统发展的轨迹，像 GPT-3 这样的系统可能具有幼儿级别的智能，而 GPT-4 则更像是聪明的高中生。在未来几年，我们将看到它们在特定任务上达到博士学位水平的智能。”主持人问到何时会出现这样的系统，穆拉提的答案是：“一年半吧。或许那时候就会有能在很多领域超越人类表现的 AI 系统了。”即 OPENAI 公司的计划是在 2025 年底或 2026 年初即将推出下一个 GPT 版本，其智商大约是博士水平。

那么 2026 年之后呢？AI 的智力水平会发展到何种地步？

按照 GPT 3（千亿参数）与 GPT 4（万亿参数）的表现，如果将参数类比生物大脑中的神经突触，即神经元之间的连接，则能够看出一些端倪。

数字神经网络在概念上类似于生物大脑。人类大脑中突触数量最常见的引用数字大约是 100 万亿（约 1000 亿神经元，每个神经元约 1000 个神经突触），这意味着当大模型参数量达到 100 万亿时，其能力可能将达到 AGI 水平。

表14：不同动物/与人的神经元、神经突触数量比较

动物	神经元数量（大约）	神经突触数量（大约）
蚯蚓 (Eisenia fetida)	300	不详
果蝇 (Drosophila melanogaster)	100,000	不详
小鼠 (Mus musculus)	70,000,000	数亿
大鼠 (Rattus norvegicus)	200,000,000	数亿
猫	250,000,000	数十亿
狗	530,000,000	数十亿
猩猩 (Pan troglodytes)	8,000,000,000	数万亿
人类 (Homo sapiens)	86,000,000,000（860 亿）	约 100 万亿到 1000 万亿

资料来源：通义千问，国信证券经济研究所整理

目前业界有专家认为基于当前人工智能技术的进步速度，AGI 可能在 2025 年至 2029 年间实现。其中：

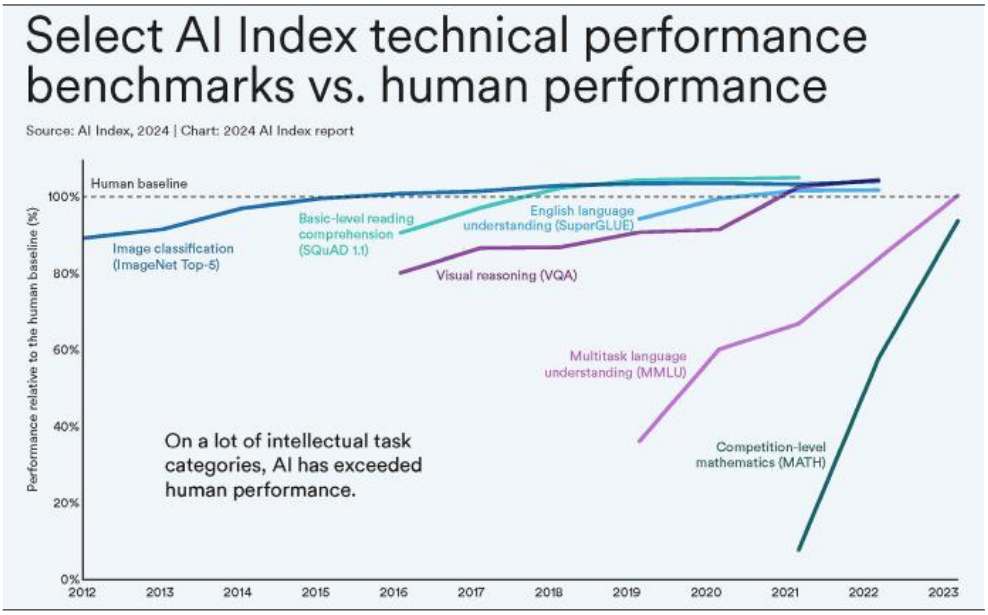
1、特斯拉 CEO 马斯克在 2024 年 3 月的观点是：“明年（2025 年），AI 可能会比任何一个人类都要聪明。到 2029 年，它可能比所有人类加起来还要聪明”；

2、英伟达 CEO 黄仁勋预计：“如果你对 AGI 的定义是通过人类的测试，那么我会告诉你，5 年就可以实现。但如果你稍微改变一下问题的提问方式，即 AGI 是要具备人类的智能，那么我还不太确定如何明确定义你们所有的智能。事实上，没有人真正确定”；

3、OPENAI 公司前员工 Leopold Aschenbrenner 认为有很大概率将在 2027 年实现 AGI 系统。他提出，算力和算法效率都在以每年 0.5 个数量级（1 个数量级=10 倍）的速度发展，再加上可能出现的释放模型性能的技术突破。（按照这个判断我们可以推算大约 1.5-2 年（2025-2026 年）模型参数将达到 10 万亿级，然后在 2027-2028 年模型参数达到 100 万亿级）。同时，Aschenbrenner 还为不同的 GPT 版本对标了人类智力：GPT-2 模型的水平大概与学龄前儿童相仿，GPT-3 模型大致达到了小学生的水平，GPT-4 实现了与较为聪明的高中生相似的水平。

即便我们认为以上推论基本合理，但我们依然猜不到 AI 具体会在哪一年完成哪些事情。AI 的单项能力一直在被证明，比如 1997 年的国际象棋，2016 年的围棋。目前来看，人工智能在多种语言理解和视觉理解基准上都超越了人类。截至 2023 年，其基础模型仍然缺乏高级推理和规划能力，但预计会取得快速进展。

图40：AI 已经在多种任务上达到人类水平



资料来源：维基百科，国信证券经济研究所整理

AGI 的五步走

OPENAI 公司将通往 AGI 之路分成了五个步骤，分别是聊天机器人、推理者、智能体、创新者、组织者。并认为，目前的 Chat GPT 4 处在第一级。

第二级推理者（Reasoners）主要是强化算法而非提升参数级别。OpenAI 待发布的 Strawberry 版本 GPT，主要将从几个方面接近这一目标：

1、采用“后训练”技术，这意味着在大量通用数据训练后，模型将被进一步调整以提升在特定任务上的性能；

2、引入多模态数据融合技术，能够将文本、图像、音频等多种类型的数据进行整合和分析，为推理提供更加全面和深入的依据。这种能力使得模型能够更好地理解和处理现实世界中的复杂情况；

3、增强 AI 模型的高级推理能力，这意味着它将能够更好地处理需要深层次理解和逻辑推断的任务；

4、减少模型产生的幻觉，即模型生成的信息虽然听起来合理但实际上并不正确的情况；

5、进一步优化传统版本的安全性与可靠性，如优化偏见、安全性和可控性等问题。

表15：OpenAI 对 AGI 路径展望

级别	名称	描述
第一级	聊天机器人（Chatbots）	具备语言对话能力的人工智能，如 ChatGPT 等。这一级别的 AI 能够与人类进行自然语言交互，回答简单问题，提供基础信息。
第二级	推理者（Reasoners）	具备人类的推理水平，能解决多种复杂难题的人工智能。这一级别的 AI 不仅能够理解语言，还能进行逻辑推理、分析复杂问题，并给出解决方案。OpenAI 认为其即将迈入这一级别，即能够解决类似博士水平的基本问题。
第三级	智能体（Agents）	能够代表用户自主采取行动，执行任务的人工智能。这一级别的 AI 不仅具备思考和推理能力，还能根据用户的指令或预设目标，自主制定计划并执行任务，如自动驾驶汽车、智能家居系统等。
第四级	创新者（Innovators）	可以协助人类完成新发明的人工智能。这一级别的 AI 不仅具备高度智能，还能在科学研究、技术创新等领域发挥重要作用，提出新的理论、设计新的产品等。
第五级	组织者（Organizations）	能够完成组织工作的人工智能。这一级别的 AI 已经具备了高度的自主性和智能性，能够像人类组织一样进行复杂的决策、规划和管理，如管理企业、协调团队等。

资料来源：搜狐，国信证券经济研究所整理

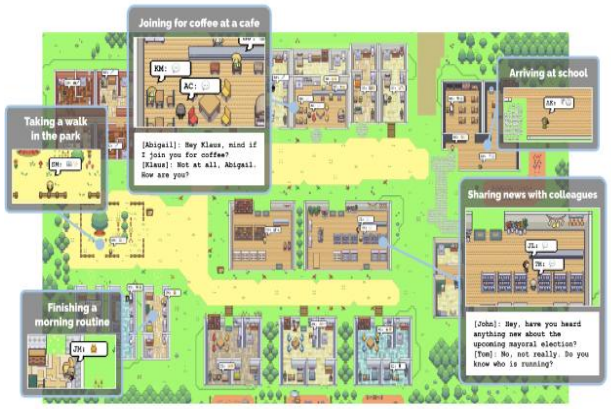
第二级很大的变化是增加了第一级的所欠缺的记忆、反思、推理、计划能力，这更像人类。因此，在第二级的基础上，则有了第三级代理（AI Agent），或者我们可以称之为智能体，即 AI 可以按照给定的目标自主完成任务。

2023 年 4 月，斯坦福大学与谷歌科研人员发表了一篇论文《生成代理：人类行为的交互式模拟》，该论文详细解释了斯坦福的“小镇”的实现机制，这个实验中包含 25 个由 AI 驱动的智能体（Agent）。这些智能体只有预设的身份和初始记忆，其所有的行为都是由 AI 驱动产生。

图41：“斯坦福小镇”实验

资料来源：斯坦福大学，国信证券经济研究所整理

图42：“斯坦福小镇”实验



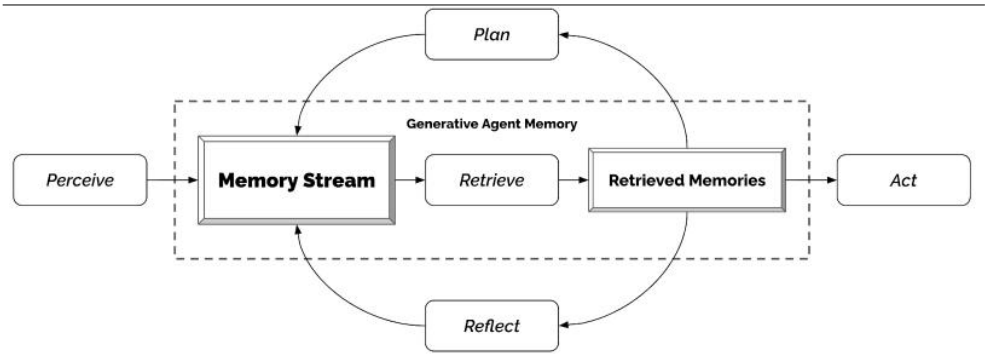
资料来源：斯坦福大学，国信证券经济研究所整理

研究人员为他们提供了一份简短的传记，包括姓名、年龄、工作、家庭、兴趣和一些习惯，然后让他们自由发挥。然后，这些 AI 居民依靠一个大型语言模型根据他们规定的传记生成他们的行为。

结果是：AI 居民行为方式与人类相似。他们醒来、做早餐、去上班、吃午餐，并与遇到的其他代理聊天。他们还会记住发生的事情，反思并制定计划。例如，当负责这个景观的研究人员建议一个角色策划一个情人节派对时，她邀请了朋友和熟人，其中许多人都在正确的时间和地点出现！

论文作者之一的 Joon Sung Park 认为，代理架构图是团队的主要技术贡献，它具有简单性：角色的感知被输入到他们的记忆流中，反馈回路允许记忆检索，这反过来又允许在代理采取行动之前进行反思和规划。Park 提及他们对架构图进行多次迭代，将其从非常复杂的想法提炼为简单而富有表现力的东西。

图43：“斯坦福小镇”智能体架构图



资料来源：斯坦福大学，国信证券经济研究所整理

研究团队发现，AI 居民制造了的社会行为。“我们没有在社会层面设计任何东西。这完全取决于智能体，”然而 AI 居民计划并参加了情人节派对；一个角色告诉其他人他正在竞选公职，他们记得这件事并互相讨论；另一个角色邀请某人约会。

当然，实验中的 AI 居民行为并不总是得体，例如：他们甚至对亲近的家人说话也非常正式；同时使用同一个宿舍厕所；去当地的酒吧而不是咖啡馆吃午饭，好像他们已经患上了白天喝酒的毛病。这些问题都可以通过未来的优化得到解决。

第四级是可以协助人类完成新发明的人工智能。这一级别的 AI 不仅具备高度智能，还能在科学研究、技术创新等领域发挥重要作用，提出新的理论、设计新的产品等。例如，谷歌 DeepMind 发布的 AlphaFold 模型，可以预测蛋白质结构，加速生物、医学研究和新药的发现效率。

第五级是能够完成组织工作的人工智能。这一级别的 AI 已经具备了高度的自主性和智能性，能够像人类组织一样进行复杂的决策、规划和管理，如管理企业、协调团队等。通俗的说，第五级是“能够创造和管理机器人的机器人”。

虽然 OPENAI 没有提及何年到第五级水平，但按照目前的进度，很有可能随着 AI 智能体的出现，多个 AI 智能体的结合价值模型自身的进化，可能在 2026 年之后，就将能力带入到第三级或者第四级。第五级实际上也是相对的，比如针对围棋、蛋白质结构预测这样比较垂类的、明晰的任务，目前已经可以做到，而相信届时第五级能力的边际将会大大拓展。

AI AGENT：下一个风口？

在斯坦福小镇实验中，很重要的概念是 AI Agent，那么如何理解 AI Agent 呢？

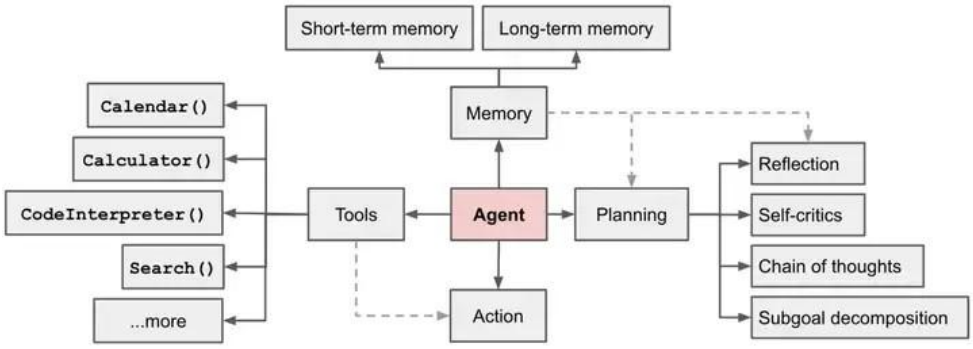
百度百科的定义是：AI Agent 是一种能够感知环境、进行决策和执行动作的智能实体。不同于传统的人工智能，AI Agent 具备通过独立思考、调用工具去逐步完成给定目标的能力。

维基百科的定义是：AI Agent 是一种能够感知环境、自主采取行动以实现目标并可能通过学习或获取知识来提高其性能的智能体。

随着大模型（LLM）的发展，今天的 AI Agent 可以理解为是一个基于大语言模型的，规划具备思考能力、记忆能力、使用工具函数的能力，能够自主完成给定任务的计算机程序。即：

AI Agent = LLM + 记忆 + 规划 + 工具

图44：AI Agent 架构图



资料来源：lmodel.net，国信证券经济研究所整理

其中：

规划（Planning）：将大型任务分割为子任务，并规划执行任务的流程；智能熟悉对任务执行的过程进行思考和反思，从而决定是继续执行任务，或判断任务完成结束并终止运行；

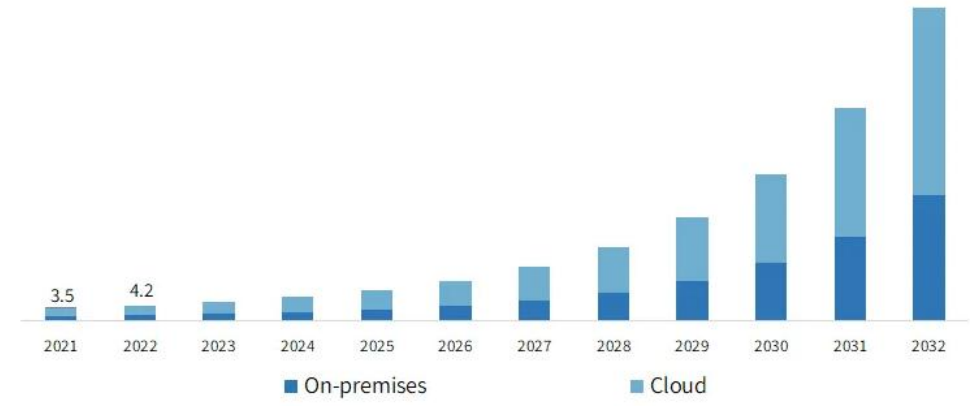
记忆（Memory）：短期记忆，是指在执行任务的过程中的上下文，会在子任务的执行过程中产生和暂存，在任务完成结束后被清空。长期记忆是指长时间保留的信息，一般是指外部知识库，通常用于提供数据库来存储和检索。更广义的说，感知也可以是记忆的一部分；

工具使用（Tool use）：为智能体配备工具 API，比如：外汇查询器、网页搜索、代码执行器、数据库查询、图片生成、文字语音转换等等。

正是由于这两年 LLM 能力的突飞猛进，使得 AI Agent 表现的能力也变得不俗。更为重要的是，目前海外、国内多达数十上百个平台提供了 AI Agent 的构建支持，在这些平台上，用户可以轻易编写 Agent，调用大模型与各类工具，以及在前人的思路上再优化，这大大降低了 AI Agent 的创作门槛。

2023 年 11 月，Gminsights 发布的报告预测，2022 年自主 AI 和自主代理市场规模价值为 42 亿美元，预计 2023 年至 2032 年期间的复合年增长率将超过 36.5%。到 2032 年，市场规模高达 881 亿美元。

图45：全球自主 AI/AI Agent 市场规模（十亿美元）



资料来源：gminsights.com，国信证券经济研究所整理

AI Agent 与 LLM 可谓是相得益彰，前者是为了具体应用目标，后者提供基础能力；前者努力探索解决问题的最佳方案，后者也可以从反馈的信息再优化自身的能力。有了五花八门的 AI Agent，LLM 也将逐渐脱离聊天范畴，而逐渐变成通用智能。

比如：一个可以设计游戏的 Agent（产品经理兼任美工），一个可以编写软件的 Agent（开发工程师），一个可以检查 bug 的 Agent（测试工程师），产品经理 Agent 可以把策划交给开发 Agent，测试 Agent 负责检查反馈问题。这样一个由 3 个 Agent 构建的游戏创作团队就完成了。理论上，随着 LLM 能力的增强，或者通过不同 LLM 的互补，以及多次配合与训练，一个自动游戏开发团队就诞生了！

依此来看，当 80 年代 IBM 5150 将行业标准确认下来之后，DOS 成为了标准，于是有了各种 PC 应用软件；

90 年代 WWW 联盟成立后，新标准来促进业界成员间的兼容性和协议的统一，涌现了诸多的网页与网站；

10 年代 iPhone、安卓手机诞生，颠覆了 PALM、WM、塞班、黑莓系统，让移动 APP 迎来了大发展时代；

20 年代的现在，基于 transformer 架构的大模型出现之后，GPT 及竞争者占据了平台地位，而 AI AGENT 将迎来繁荣期。

表16：AI AGENT 类比历史不同科技时代的地位

时代	时间	标准	平台	应用
计算机	20 世纪 80 年代	IBM 兼容机	DOS	应用软件
互联网	20 世纪 90 年代	WWW 联盟	网景、IE 浏览器	网页与网站
移动互联网	21 世纪 10 年代	iPhone、安卓手机	IOS、安卓系统	移动 APP
通用人工智能	21 世纪 20 年代	大模型	GPT 及竞争者	AI AGENT

资料来源：国信证券经济研究所整理

并非只有平台才有机会：

在计算机时代，应用软件走出了甲骨文、Adobe 这样的巨头；

在互联网时代，网页与网站走出了亚马逊、谷歌、脸书这样的巨头；

在移动互联网时代，APP 走出了微信、Whats-app 这样的超级应用；

我们期待着在 AI AGENT 中，“千淘万漉虽辛苦，吹尽黄沙始到金”，能同样将走出超级 AGENT。

这或将是比互联网、移动互联网，更加火热的时代！

小结

从 2016 年英特尔放弃“tick-tock”开始，业界就开始呼唤一种新路线来完成对算力升级的接力。GPU 通过其并行、通用运算的特征，自然而然地接过了 CPU 的接力棒。在 10 年的时间里，英伟达单个 GPU 在 AI 推理方面的性能大幅提升了 1000 倍，这甚至超越了摩尔定律的期望，给后来的算力应用奠定了坚实的基础。

算力应用场景不同，但可初步归纳为三个方向：加密币的挖矿需求，云计算的建设需求，以及新能源车的车载智能化需求。

某种意义上，云计算的出现是因为其“弹性”的本质：即多余的算力可以租赁给需要的人，这样可以实现资源效率的最大化。在 2006 年亚马逊 EC2 发布之后的十年时间里，云计算与需求之间始终以一种相对平衡、稳定的方式交互增长。人们也没有想到有一种特别的场景会打破这种平衡。

直到 2017 年 Transformer 架构的出现，事情开始起变化。Transformer 架构与神经网络架构不同，它并非像 CNN 那样将运算步骤分层而将“复杂问题简单化”，而是将所有的数据都呈现在一个巨大的多维向量空间中，通过计算其向量之间的距离而得到概率，再通过加权汇总问题的概率而形成计算机意义上的“理解”，这个算法被形象地称为“大力出奇迹”（Scaling law）。

在最初的几年，Transformer 架构生成的 BERT 模型与 GPT 早期模型的参数量仅在 1 亿或者 10 亿级别，模型的效果一般，因此并未引发科技界的广泛关注。直到千亿参数 GPT3.0 出现之后，其在大语言上的表现颠覆了以往任何一个人工智能模型，终于征服了科技界，让人类见识到“大力出奇迹”的效果居然可以如此惊艳——原来之前没有出现“奇迹”是因为还不够“大力”！因此，2022 年的 GPT 3.0/3.5 的出现，好比打开了通往 AGI 的潘多拉魔盒，将全世界的注意力都吸引了过来。

如果千亿参数类比小学生，万亿参数是高中生，那么十万亿参数是否是博士生甚至更高水平？百万亿参数是否可以成为人类千百万年来苦苦追寻的 AGI 的开始？全世界都太想尽早揭开这个谜底，尽管在未来几年，将参数扩大 100 倍代表着更加巨大的投资，但这如同我们在茫茫大海中已经眺望到了远方模糊的新大陆，或者在黑暗的山洞里已经远远看到了宝藏的光芒！

尽管降低 GPU 的成本，优化模型的架构，增加电力的供给...都可能是接下来面临的现实的困难。但不能忽视的是，随着人类越接近那个目标，还将面临更大的挑战——我们走得过快，不是么？人类准备好有关隐私、安全、伦理、价值观等层面的挑战了么？在百年未有之大变局中，以上问题在今天的世界中都被处理的囫囵囵囵，这是一句轻描淡写的“对齐”就能够解决的问题吗？如同一个不够孝顺的家长，期望孩子能够做到仁义礼智，此间的“对齐”由何而来？

因此，对 AGI 的期望不要过高，因为我们可能缺乏系统性的驾驭它们的能力，尤其是价值观上的共识，这或许是下一个十年、几十年都不得不去面临的更长期的问题；而对 AGI 的期望也不能过低，因为它是过去 60 年中对计算机、互联网、移动互联网、云计算的一个系统性的完结，它的光芒也必然超过以往的任何时代。

不论此刻我们怀着期待、憧憬、兴奋，还是不安、担忧的心情，科技不朽的车轮终将前进，我们也只能带着发展的眼光在前进中不断学习、提升自身，迎接挑战！

附录：本时期重大事件

表17：2016-2024 大事记

年份	重大事件
2016 年	英伟达 PASCAL 架构，深度学习计算机 DGX-1，Drive PX2，谷歌 Alphago 战胜李世石，微软 HoloLens ，微软收购 LinkedIn，墨子号量子通信卫星，寒武纪成立，英特尔放弃 tick-tock 战略。
2017 年	英伟达发布 Volta 架构，微软发布 Xbox One X，特斯拉发布 Model 3，Transformer 架构被提出。
2018 年	英伟达 Turing 架构，Opensea 成立，贝壳成立，Sky Mavis 发布区块链游戏 Axie Infinity。
2019 年	星链卫星发射，谷歌 54 量子位计算机 Sycamore，英伟达 Ampere 架构，瑞典商用无人驾驶卡车 T-pod，中国 5G 商用，嫦娥四号探测器成功登陆月球背面。
2020 年	特斯拉 Model Y，沐熙集成成立，谷歌发布基于 transformer 架构的 BERT 模型。
2021 年	微软 Windows 11，脸书改名 Meta，OPENAI 发布 DALL-E，北汽极狐阿尔法 S（华为智驾 HI 版）发布。
2022 年	OPENAI 发布 ChatGPT 3.5，英伟达 Hopper 架构，Midjourney 发布，OPENAI 发布 DALL-E 2，英伟达发布 H100。
2023 年	OPENAI 发布 ChatGPT 4.0，OPENAI 发布 DALL-E 3，微软发布 CoPilot，脸书发布 Llama，谷歌发布 Gemini，Stable Diffusion 发布，Tesla 发布无人驾驶 HW4.0，新能源汽车在中国销售占比超过 30%，EAST 实现稳态高约束模式等离子体运行 403 秒，“中国环流三号”实现 100 万安培等离子体电流下的高约束模式运行，百度发布文心一言，阿里巴巴发布通义千问，讯飞发布星火大模型，华为发布盘古，腾讯发布腾元，华为发布 MATE 60 手机，华为发布问界 M7、问界 M9，北斗定位服务日均使用量已超过 6000 亿次。
2024 年	中国量子计算机本源悟空上线，OPENAI 推出 SORA，英伟达发布 GB200，中国嫦娥六号采集月球背面月壤，百度在武汉商用 Robot taxi，华为发布三折屏手机 MateXT、原生鸿蒙系统。

资料来源：各公司网站，国信证券经济研究所整理

风险提示

地缘政治的不确定性，海外降息幅度的不确定性，部分行业竞争格局的不确定性。

免责声明

分析师声明

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后 6 到 12 个月内的相对市场表现，也即报告发布日后的 6 到 12 个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A 股市场以沪深 300 指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数（HSI.HI）作为基准；美国市场以标普 500 指数（SPX.GI）或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数 10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数 10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数 10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数 10%以上

重要声明

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券经济研究所

深圳
深圳市福田区福华一路 125 号国信金融大厦 36 层
邮编：518046　总机：0755-82130833

上海
上海浦东民生路 1199 弄证大五道口广场 1 号楼 12 层
邮编：200135

北京
北京西城区金融大街兴盛街 6 号国信证券 9 层
邮编：100032