主题研究

AI 芯片：应用落地推动产品多样化

观点聚焦

分析员 黄乐平
SAC 执证编号：S0080518070001
SFC CE Ref：AUZ066
leping.huang@cicc.com.cn

分析员 何玫，CFA
SAC 执证编号：S0080512090005
SFC CE Ref：AVJ148
mei.he@ciccus.com

联系人 杨俊杰
SAC 执证编号：S0080117090047
junjie.yang@cicc.com.cn

投资建议

过去一年，随着 AI 在各个行业的不断落地，行业痛点逐渐被发现，AI 芯片的发展路径渐趋明朗。未来五年，我们预计 AI 芯片市场的规模有接近 10 倍的增长，2022 年将达到 352 亿美元。在训练芯片方面，我们继续看好 Nvidia 利用 CUDA+GPU 构建的生态环境优势，以 Google TPU 为代表的 xPU 很难撼动其垄断地位。随着终端细分场景落地，推断芯片的发展出现专业化趋势，为现有及初创公司提供巨大发展空间。我们预计，推断芯片市场规模到 2022 年会超过训练芯片。中国 AI 芯片设计企业中，建议关注海思、展锐，比特大陆，寒武纪，地平线，异构智能，云天励飞，龙加智。产业链上，建议关注台积电，及 IP 厂商 Synopsys、Cadence、GUC 等。

理由

Nvidia 在训练芯片上的垄断地位短期很难被撼动：过去一年，为维持其训练芯片市场的王者地位，Nvidia 推出 Volta 架构以提升 GPU 性能，并更新 CUDA 异构计算平台继续巩固其开发者生态环境。一年前，我们担心的“Google 会影响 NVIDIA 在 AI 芯片的领导地位吗？”情况并未出现。我们预计训练芯片市场未来五年将保持 54%增速，到 2022 年达到 172 亿美金规模。

推断芯片开始专业化分工：在手机芯片方面，苹果、高通、海思、联发科等芯片公司相继推出支持 AI 加速功能的新一代芯片，实现产品附加价值的提升。寒武纪等初创公司及 ARM，Cadence 等设计企业主要通过 IP 授权方式为行业提供附加价值。安防芯片方面，海思、安霸等传统视频解码芯片厂商也推出了支持 AI 加速的新一代产品，与 Nvidia，Mobileye 的终端推断芯片形成有力竞争。

看好初创企业在云端推断和无人驾驶领域的发展机会：AI 在搜索，电商，语音交互上的大规模应用，为 AI 云端推断芯片提供了广阔的市场。由于各种场景对功耗、响应速度等要求不同，硬件针对不同算法需要做特殊优化，这也为独立芯片设计公司提供了广阔的发展前景。自动驾驶 AI 芯片上，虽然 Mobileye、Nvidia 暂时领先，但因真正的全自动驾驶实现起来非常复杂，存在不同的发展路径，对 Tesla 等整车企业以及初创公司来说都有发展机会。

AI 芯片发展对产业链的影响：目前 AI 芯片设计百花齐放的格局，将使 Synopsys、Cadence 等 IP 授权商，及 GUC 等 IC 后端设计商受益。此外，由于 AI 芯片一般采用 10nm 以上先进制程，目前利好代工厂商台积电。我们也应长期关注中芯国际在先进工艺上的发展进度。

风险

人工智能落地速度不及预期。

相关研究报告

- 科技 | AI+安防：AI 助力安防行业拓宽边界 (2018.08.16)
- 主题研究 | AI+消费电子：谁偷走了我们的时间？ (2018.08.15)
- 主题研究 | AI+智能家居：小米、格力谁能成为 IoT 时代的王者 (2018.08.15)
- 科技,可选消费 | AI+零售：不只是无人零售，看人工智能如何撬动零售变革 (2018.08.14)
- 金融,科技 | AI+金融：人工智能成为长远发展的有力发动机 (2018.08.13)
- 主题研究 | 人工智能芯片的中国突围 (2018.03.14)
- AI+半导体，GOOGLE 会影响 NVIDIA 在 AI 芯片的领导地位吗？(2017.04.10)
- 主题研究 | 寻找 AI+淘金热中的卖水人 (2016.12.04)

目录

图表

AI 芯片：场景渐趋明朗，呈现专业化发展

新品竞相发布，AI 芯片行业格局渐趋明朗

AI 芯片设计是人工智能产业链的重要一环。自 2017 年 5 月以来，各 AI 芯片厂商的新品竞相发布，经过一年多的发展，各环节分工逐渐明显。AI 芯片的应用场景不再局限于云端，部署于智能手机、安防摄像头、及自动驾驶汽车等终端的各项产品日趋丰富。除了追求性能提升外，AI 芯片也逐渐专注于特殊场景的优化。

图表 1：自 2017 年 5 月以来发布的 AI 芯片一览

时间	企业	产品类型	具体内容
2017年5月	Nvidia	云端芯片	发布最新GPU Volta 架构芯片
2017年5月	Google	云端芯片	发布TPU 2.0
2017年5月	ARM	智能手机芯片相关技术	发布针对AI优化的DynamicIQ芯片架构
2017年8月	Intel	安防/无人机芯片	推出新的Movidius Myriad X VPU
2017年8月	百度	云端芯片	发布XPU，一款256核基于FPGA的云计算加速芯片
2017年9月	Intel	云端芯片	推出自学习神经元芯片Loihi，采用14nm工艺
2017年9月	华为海思	智能手机芯片	发布人工智能芯片“Kirin 970”
2017年10月	Apple	智能手机芯片	发布iPhone X，首次使用A11 Bionic芯片，搭载神经网络引擎
2017年10月	深鉴科技	安防芯片相关技术	发布人脸识别模组、ARISTOTLE架构平台等
2017年11月	寒武纪	智能手机芯片IP	发布 Cambricon 1H8/1H16/1M芯片
2017年11月	比特大陆	云端芯片	发布全球首款张量加速计算芯片BM1680等
2017年12月	Qualcomm	智能手机芯片	发布Snapdragon 845移动平台，采用10nm工艺，支持多种深度学习框架
2017年12月	地平线机器人	安防/自动驾驶芯片	发布“旭日”和“征程”两款嵌入式AI芯片，分别面向智能驾驶和智能摄像头
2018年1月	Nvidia	自动驾驶芯片	发布用于自动驾驶的Jetson Xavier芯片，及车载计算机Drive PX Pegasus，搭载两块Xavier SoC，算力完全支持L5
2018年1月	异构智能	云端芯片	发布NovuTensor一代 AI芯片
2018年4月	地平线机器人	自动驾驶芯片	发布“征程2.0”芯片及MATRIX 1.0自动驾驶计算平台
2018年5月	Google	云端芯片	发布TPU 3.0
2018年5月	寒武纪	云端芯片	发布MLU 100云端智能芯片

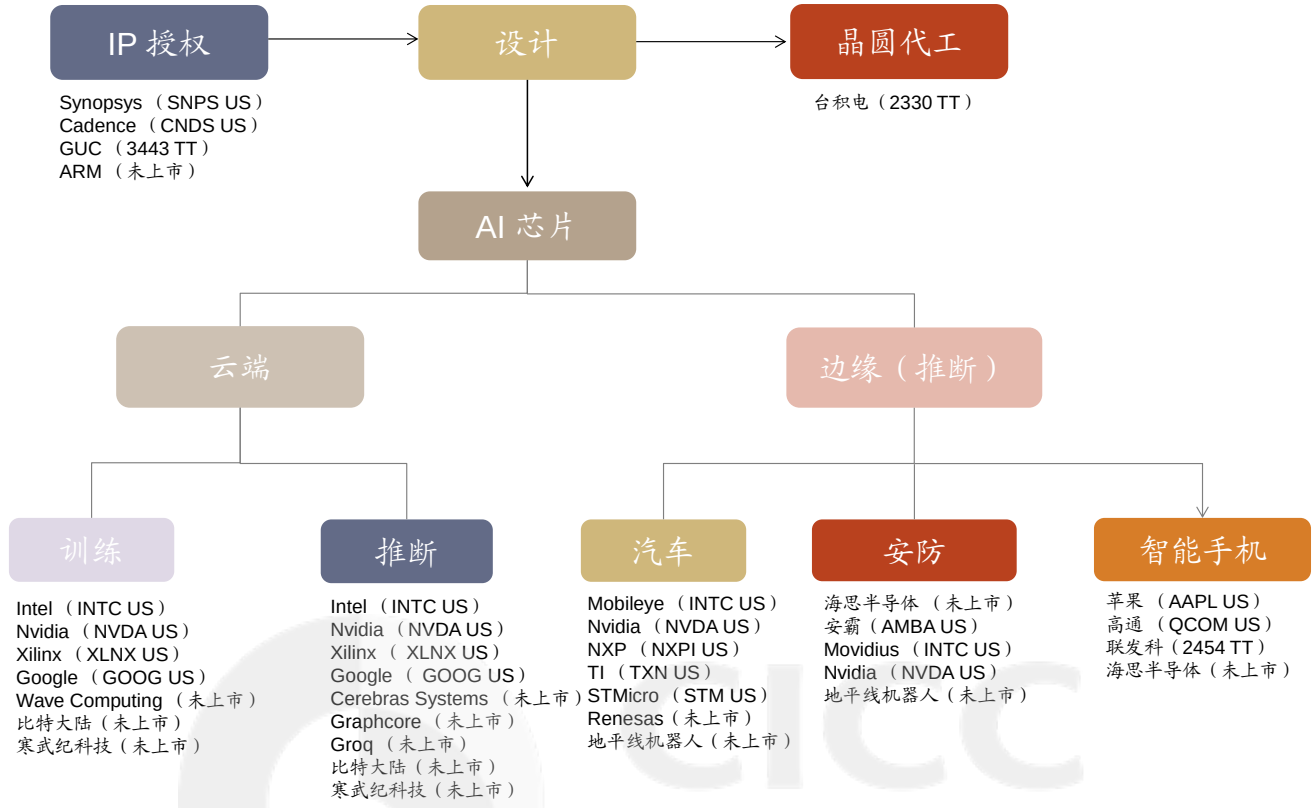
资料来源：公司网站，中金公司研究部

目前，人工智能产业链中，包括提供 AI 加速核的 IP 授权商，各种 AI 芯片设计公司，以及晶圆代工企业。

- 按部署的位置来分，AI 芯片可以部署在数据中心（云端），和手机，安防摄像头，汽车等终端上。
- 按承担的任务来分，可以被分为用于构建神经网络模型的训练芯片，与利用神经网络模型进行推断的推断芯片。训练芯片注重绝对的计算能力，而推断芯片更注重综合指标，单位能耗算力、时延、成本等都要考虑。
- 训练芯片受算力约束，一般只在云端部署。推断芯片按照不同应用场景，分为手机边缘推断芯片、安防边缘推断芯片、自动驾驶边缘推断芯片。为方便起见，我们也称它们为手机 AI 芯片、安防 AI 芯片和汽车 AI 芯片。

- 由于 AI 芯片对单位能耗算力要求较高，一般采用 14nm/12nm/10nm 等先进工艺生产。台积电目前和 Nvidia、Xilinx 等多家芯片厂商展开合作，攻坚 7nm AI 芯片。

图表 2: AI 芯片投资地图



资料来源：中金公司研究部

图表 3: AI 芯片产业链主要公司估值表

Ticker	Company	CICC Rating	M/Cap USD mn	Price 22-Aug	Target Price	P/E 2018E	P/E 2019E	P/B 2018E	P/B 2019E	ROE(%) 2018E	EPS Growth 2018E	1D	5D	1M	YTD
2330 TT	台积电 (TSMC)	NA	212,059	241.00	NA	17.5	15.5	3.7	3.4	22.2	3%	1	-1	1	5
INTC US	英特尔 (INTEL)	NA	219,576	47.62	NA	11.5	11.2	2.9	2.5	22.2	99%	2	-1	-8	5
NVDA US	英伟达 (NVIDIA)	NA	154,019	253.32	NA	56.7	32.0	21.8	13.5	40.9	-11%	2	-3	1	31
XLNX US	赛灵思 (XILINX)	NA	18,316	72.42	NA	28.7	22.5	7.7	7.5	22.5	24%	2	2	7	9
3443 TT	创意电子 (GUC)	NA	1,281	293.00	NA	33.0	24.5	8.7	7.3	26.6	39%	4	-2	-9	15
CDNS US	铿腾电子 (Cadence)	NA	12,839	45.40	NA	27.0	24.7	9.7	7.4	40.9	133%	1	-1	1	9
SNPS US	新思科技 (Synopsys)	NA	13,995	93.92	NA	24.7	22.6	4.1	4.0	17.3	316%	0	1	2	10

资料来源：万得资讯，彭博资讯，中金公司研究部

AI 芯片市场规模：未来五年有接近 10 倍的增长，2022 年将达到 352 亿美元

根据我们对相关上市 AI 芯片公司的收入统计，及对 AI 在各场景中渗透率的估算，2017 年 AI 芯片市场规模已达到 39.1 亿美元，具体情况如下：

- 2017 年全球数据中心 AI 芯片规模合计 23.6 亿美元，其中云端训练芯片市场规模 20.2 亿美元，云端推断芯片 3.4 亿美元。
- 2017 年全球手机 AI 芯片市场规模 3.7 亿美元。
- 2017 年全球安防摄像头 AI 芯片市场规模 3.3 亿美元。
- 2017 年全球自动驾驶 AI 芯片的市场规模在 8.5 亿美元。

图表 4: AI 芯片市场规模及竞争格局

应用场景		市场规模 2017（百万美元）	市场规模 2022（百万美元）	CAGR (2017-2022)	领导者	挑战者
云端	训练	2,015	17,212	54%	Nvidia	Google/Intel/AMD/初创公司（机会较小）
	推断	343	7,186	84%	Nvidia	Google/Intel/AMD/初创公司（有机会）
边缘	智能手机	368	3,793	59%	苹果、三星、海思、高通、联发科、展锐	初创公司（IP授权模式可能有机会）
	安防	330	1,822	41%	海思、安霸、Intel(Movidius)、Nvidia	初创公司（机会较小）
	汽车	854	5,204	44%	Intel (Mobileye)、Nvidia	初创公司（有机会）
	合计	3910	35217	55%		

资料来源：中金公司研究部

Nvidia 在 2017 年时指出，到 2020 年，全球云端训练芯片的市场规模将达到 110 亿美元，而推断芯片（云端+边缘）的市场规模将达到 150 亿美元。Intel 在刚刚结束的 2018 DCI 峰会上，也重申了数据业务驱动硬件市场增长的观点。Intel 将 2022 年与用于数据中心执行 AI 加速的 FPGA 的 TAM 预测，由 70 亿美元调高至 80 亿美元。

而同时我们也注意到：

1）手机 SoC 价格不断上升、AI 向中端机型渗透都将为行业创造更广阔的市场空间。

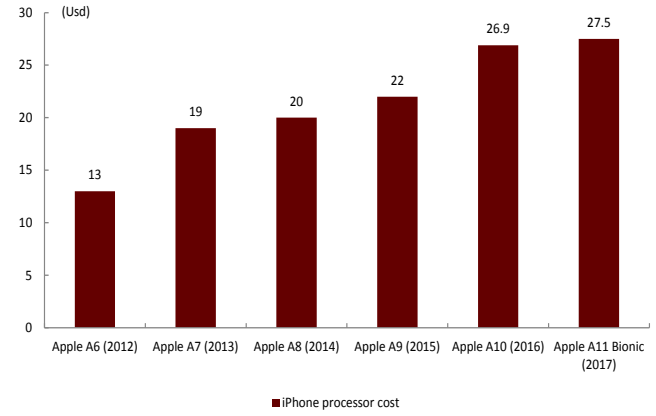
2）安防芯片受益于现有设备的智能化升级，芯片需求扩大。

3）自动驾驶方面，针对丰田公司提出的算力需求，我们看到当下芯片算力与 L5 级自动驾驶还有较大差距。英飞凌公司给出了各自动驾驶等级中的半导体价值预测，可以为我们的 TAM 估算提供参考。

结合以上观点，及我们对 AI 在各应用场景下渗透率的分析，我们预测：

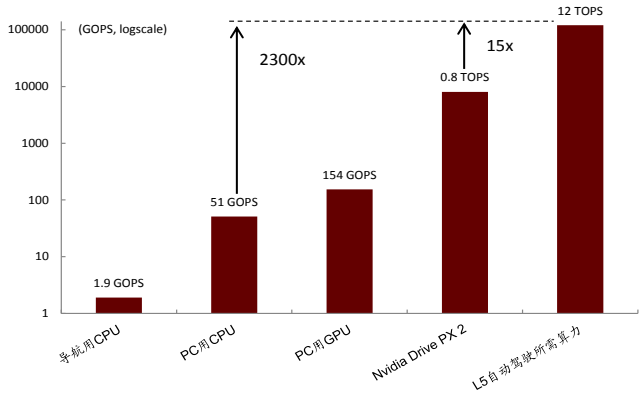
- 云端训练芯片市场规模在 2022 年将达到 172 亿美元，CAGR~54%。
- 云端推断芯片市场规模在 2022 年将达到 72 亿美元，CAGR~84%。
- 用于智能手机的边缘推断芯片市场规模 2022 年将达到 38 亿美元，CAGR~59%。
- 用于安防摄像头的边缘推断芯片市场规模 2022 年将达到 18 亿美元，CAGR~41%。
- 用于自动驾驶汽车的边缘推断芯片市场规模 2022 年将达到 52 亿美元，CAGR~44%。

图表 5：历代 Apple 手机芯片成本趋势



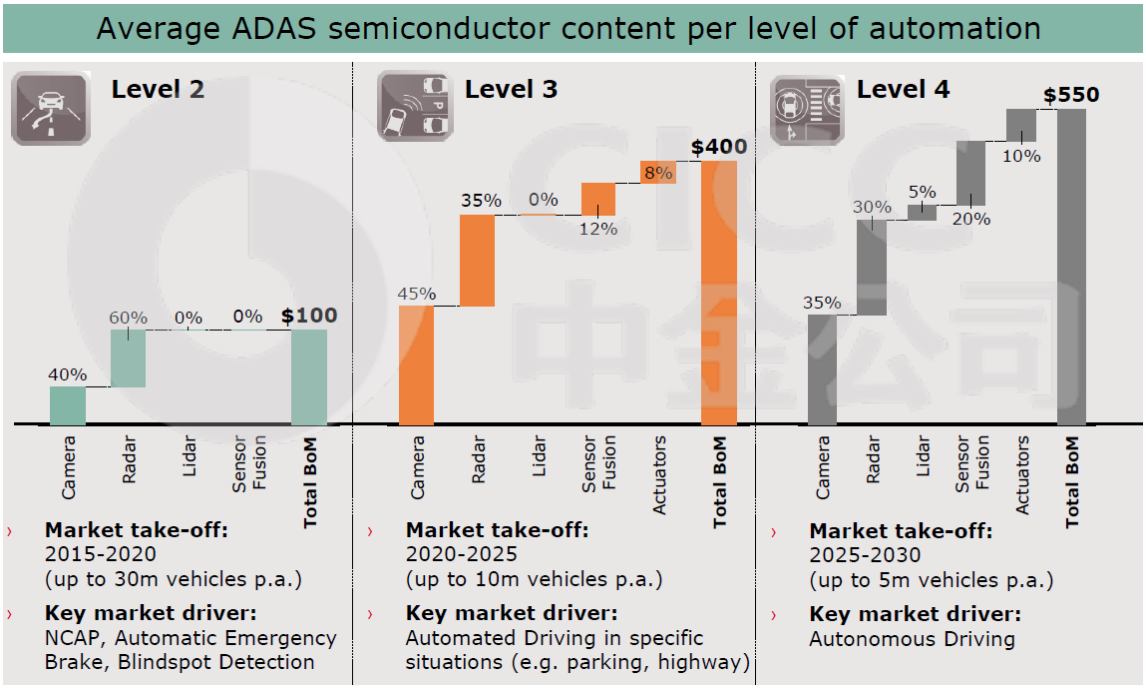
资料来源：iHS，中金公司研究部

图表 6：自动驾驶算力需求加速芯片升级



资料来源：Toyota，中金公司研究部

图表 7：英飞凌对各自动驾驶等级中半导体价值的预测



资料来源：Infineon，中金公司研究部

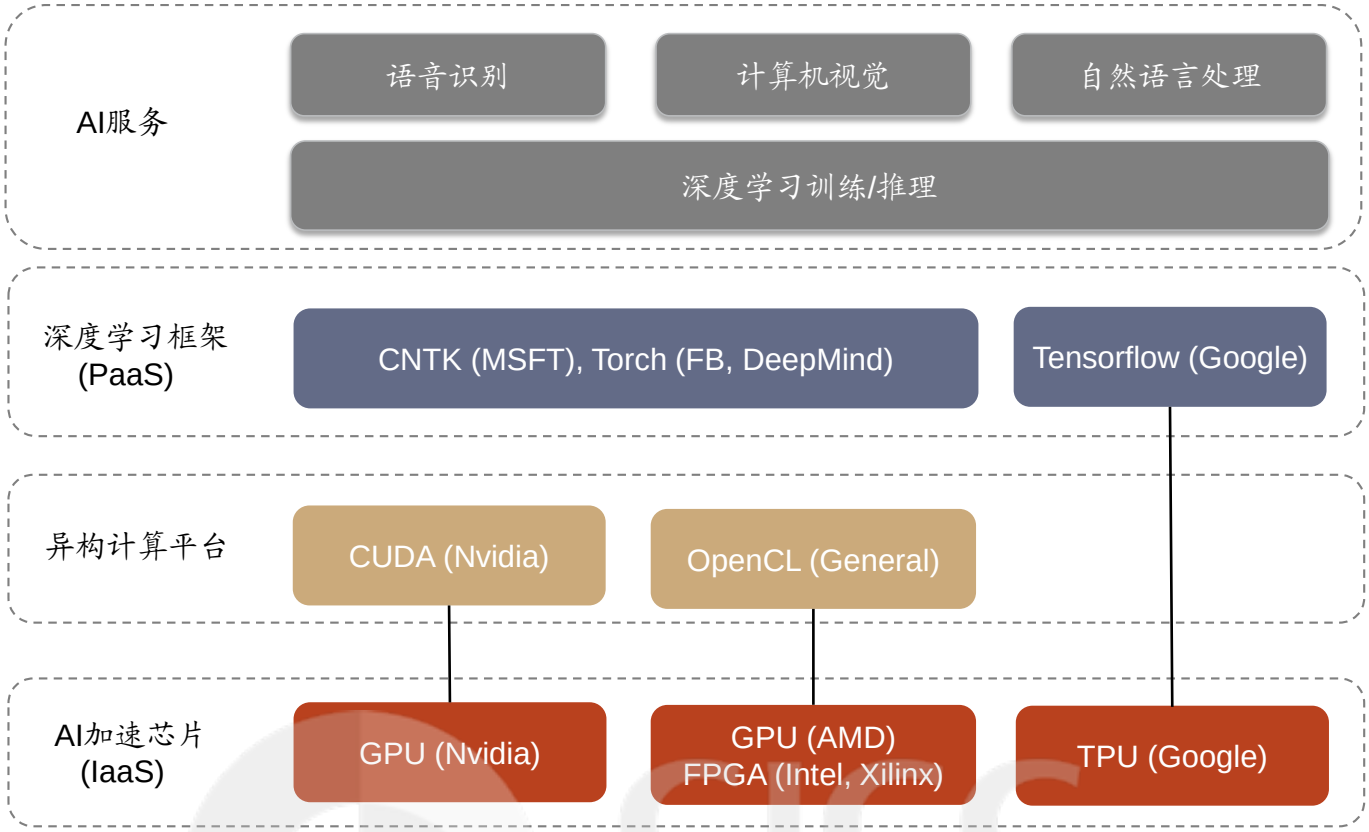
云端训练芯片：TPU 很难撼动 Nvidia GPU 的垄断地位

训练是指通过大量的数据样本，代入神经网络模型运算并反复迭代，来获得各神经元“正确”权重参数的过程。CPU 由于计算单元少，并行计算能力较弱，不适合直接执行训练任务，因此训练一般采用“CPU+加速芯片”的异构计算模式。目前 Nvidia 的 GPU+CUDA 计算平台是最成熟的 AI 训练方案，除此还有：

- 第三方异构计算平台 OpenCL + AMD GPU 或 OpenCL+Intel/Xilinx 的 FPGA。
- 云计算服务商自研加速芯片（如 Google 的 TPU）这两种方案。

各芯片厂商基于不同方案，都推出了针对于云端训练的 AI 芯片。

图表 8：AI 芯片工作流程



资料来源：中金公司研究部

图表 9：云端训练芯片对比

处理器名称	Nvidia P100 (Pascal) 第五代GPU	Nvidia V100 PCIe (Volta) 第六代GPU	Google TPU 2.0	Google TPU 3.0	Intel Stratix 10 (FPGA)	Xilinx Virtex Ultrascale+ (FPGA)
逻辑核心数	3,584（CUDA 核心）	5,120（CUDA 核心）	多核心	多核心	多核心	多核心
深度学习计算能力	10	120	45	90	最高达 10	最高达 28
缓存	4MB L2	6MB L2	NA	NA	1MB L2	1MB L2
内存大小	16GB	16GB	16GB	32GB	NA	up to 8GB
内存带宽	720GB/s	900GB/s	600GB/s	NA	最高达 512GB/s	NA
功耗	250W	250W	约 200-250W	约 200W	低	低

资料来源：Intel，Nvidia，Google，Xilinx，中金公司研究部

我们认为，从整个云端训练芯片的市场竞争格局来看，目前 Nvidia GPU 的优势暂时明显。具体情况如下：

► Nvidia

Nvidia GPU 在云端训练芯片中占据领导者地位。GPU 最初只服务于图形处理加速，为了使 GPU 能够更好地用于通用计算，Nvidia 开发了 CUDA 计算平台。CUDA 对各种主流学习框架的兼容性最好，成为 Nvidia 的核心竞争力之一。目前 Nvidia GPU 已发展到第六代 Volta 架构，5120 个 CUDA 核心提供了超 120 TFLOPS 深度学习算力，带宽高达 900GB/s，以其优异的性能继续在全球领先。

Nvidia 来自数据中心的收入从 16 财年 4 季度起开始飙升，从 9700 万美元暴增至 19 财年 2 季度的 7.6 亿美元，至今仍保持着高于 70%的同比增速，成为训练芯片中绝对的王者。目前，Nvidia V100 GPU 及次新款产品 P100 GPU 在 AWS 云、微软云、百度云中都被广泛应用。即便是 Google 的一些深度学习训练任务，同样离不开 Nvidia GPU。

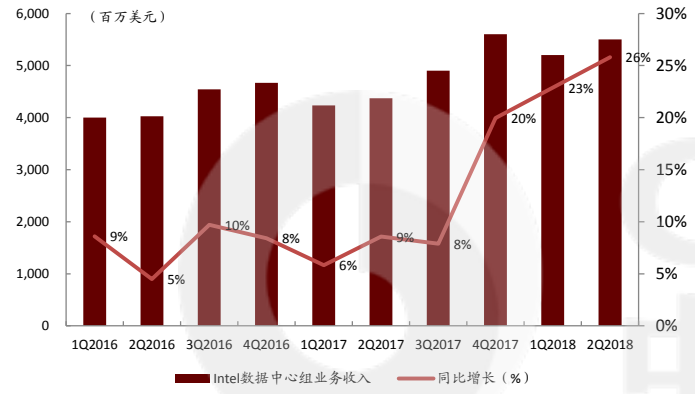
► Google

在 GPU 之外，云端训练的新入竞争者是 TPU。Google 在去年正式发布了其 TPU 芯片，并在二代产品中开始提供对训练的支持，但比较下来，GPU 仍然拥有最强大的带宽（900GB/s，保证数据吞吐量）和极高的深度学习计算能力（120 TFLOPS vs. TPUv2 45 TFLOPS），在功耗上也并没有太大劣势（TPU 进行训练时，引入浮点数计算，需要逾 200W 的功耗，远不及推断操作节能）。目前 TPU 只提供按时长付费使用的方式，并不对外直接销售，市占率暂时也难以和 Nvidia GPU 匹敌。

► Intel

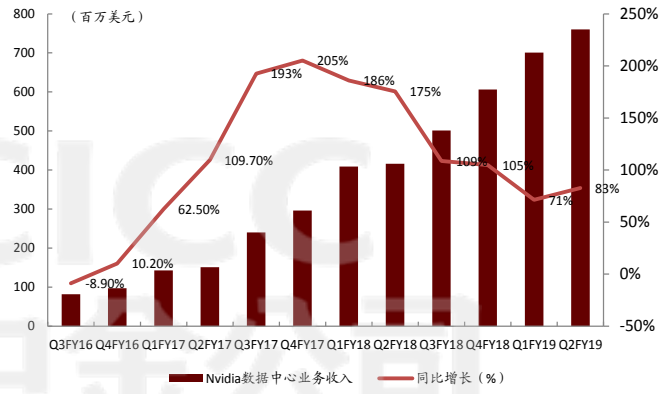
虽然深度学习任务主要由 GPU 承担，但 CPU 目前仍是云计算芯片的主体。Intel 于 2015 年底年收购全球第二大 FPGA 厂商 Altera 以后，也积极布局 CPU+FPGA 异构计算助力 AI，并持续优化 Xeon CPU 结构。2017 年 Intel 发布了用于 Xeon 服务器的，新一代标准化的加速卡，使用户可以 AI 领域进行定制计算加速。得益于庞大的云计算市场支撑，Intel 数据中心组业务收入规模一直位于全球首位，2016-17 年单季保持同比中高个位数增长。2017 年 4 季度起，收入同比增速开始爬坡至 20%左右，但相比 Nvidia 的强劲增长态势仍有差距。

图表 10: Intel 单季度数据中心组业务收入



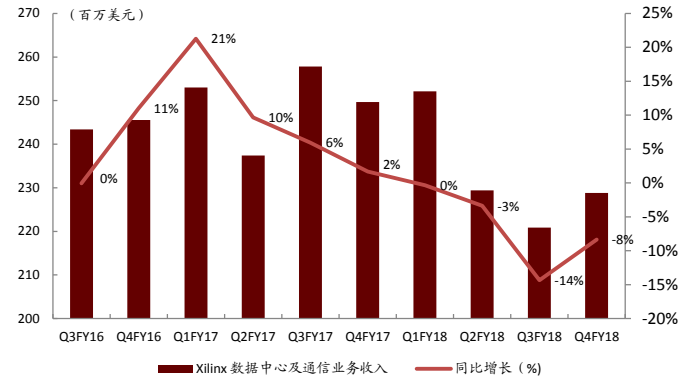
资料来源：公司季报，中金公司研究部

图表 11: Nvidia 单季度数据中心业务收入



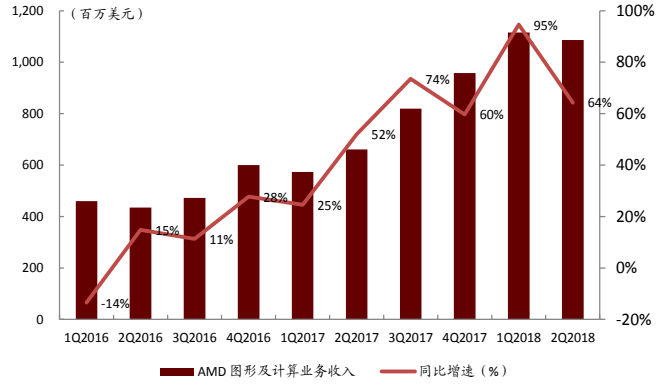
资料来源：公司季报，中金公司研究部

图表 12: Xilinx 单季度通讯&数据中心业务收入



资料来源：公司季报，中金公司研究部

图表 13: AMD 单季度计算&图形业务收入



资料来源：公司季报，中金公司研究部

► Xilinx:

Xilinx 是全球最大 FPGA 厂商，FPGA 用于深度学习训练拥有功耗上的优势，但缺点是其编程模型过于复杂，需要工程师有很强的硬件专业知识，自 18 财年 2 季度起，受 4G 资本开支下滑影响，通讯及数据中心业务收入出现同比负增长。

► AMD

AMD 虽未单独拆分数据中心收入，但从其计算和图像业务的收入增长情况来看，GPU 销量向好。目前 AMD GPU 也开始切入深度学习训练任务，但市场规模落后于 Nvidia。

云端推断芯片：百家争鸣，各有千秋

推断是指借助现有神经网络模型进行运算，利用新的输入数据来一次性获得正确结论的过程。推断过程对响应速度一般有较高要求，因此会采用 AI 芯片（搭载训练完成的神经网络模型）进行加速。

相比训练芯片，推断芯片考虑的因素更加综合：单位功耗算力，时延，成本等等。初期推断也采用 GPU 进行加速，但由于应用场景的特殊性，依据具体神经网络算法优化会带来更高的效率，FPGA/ASIC 的表现可能更突出。除了 Nvidia、Google、Xilinx、Altera（Intel）等传统芯片大厂涉足云端推断芯片以外，Wave computing、Groq 等初创公司也加入竞争。中国公司里，寒武纪、比特大陆同样积极布局云端芯片业务。

图表 14：主要云端推断芯片对比

	Google TPU 1.0	Nvidia P40 GPU	Nvidia P4 GPU	Wave computing	Groq	Cambricon MLU 100	Bitmain BM 1680
训练计算性能 (TFLOPS)	NA	12 (FP32)	5.5 (FP32)	支持训练	支持训练	支持训练	2 (FP16)
推断计算性能 (TOPS)	90 INT8	48 INT8	22 INT8	180 INT8	400 INT8	128 INT8	支持推断
片上内存	24MB	11MB	NA	NA	NA	NA	NA
功耗	75W	250W	75W	NA	50W	80W	25W
带宽	34 GB/s	350 GB/s	192GB/s	270GB/s	NA	NA	50GB/s

资料来源：Intel，Nvidia，Google，Wave computing，Groq，寒武纪科技，比特大陆，中金公司研究部

我们认为，**云端推断芯片在未来会呈现百花齐放的态势。**具体情况如下：

► Nvidia

在云端推断芯片领域，Nvidia 主打产品为 P40 和 P4，二者均采用 TSMC 16nm 制程。Tesla P4 拥有 2560 个流处理器，每秒可进行 22 万亿次（TOPS）计算（对应 INT 8）。而性能更强的 Tesla P40 拥有 3840 个流处理器，每秒可进行 47 万亿次（TOPS）计算（对应 INT 8）。从单位功耗推断能力来看，P4/P40 虽然有进步，但仍逊于 TPU。**GPU 在推断上的优势是带宽。**

► Google

Google TPU 1.0 为云端推断而生，其运算单元对神经网络中的乘加运算进行了优化，并采用整数运算。TPU 1.0 单位功耗算力在量产云端推端芯片中最强，达 1.2TOPS/Watt，优于主流 Nvidia GPU。TPU 2.0 在推断表现上相比于 1 代并没有本质提升，主要进步是引入对浮点数运算的支持，及更高的片上内存。正如前文所述，支持训练的 TPU 功耗也会变得更高。

► Wave Computing

Wave computing 于 2010 年 12 月成立于加州，目前累计融资 1.2 亿美元，是专注于云端深度学习训练和推理的初创公司。Wave computing 的一代 DPU 深度学习算力达 180 TOPS，且无需 CPU 来管理工作流。目前公司正与 Broadcomm 合作在开发二代芯片，将采用 7nm 制程。

► Groq

Groq 是由 Google TPU 初始团队离职创建的 AI 芯片公司，计划在 2018 年发布第一代 AI 芯片产品，对标英伟达的 GPU。其算力可达 400 TOPs（INT 8），单位能耗效率表现抢眼。

► 寒武纪科技

寒武纪在 2017 年 11 月发布云端芯片 MLU 100，同时支持训练和推断，但更侧重于推断。MLU 100 在 80W 的功耗下就可以达到 128 TOPS（对应 INT 8）的运算能力。

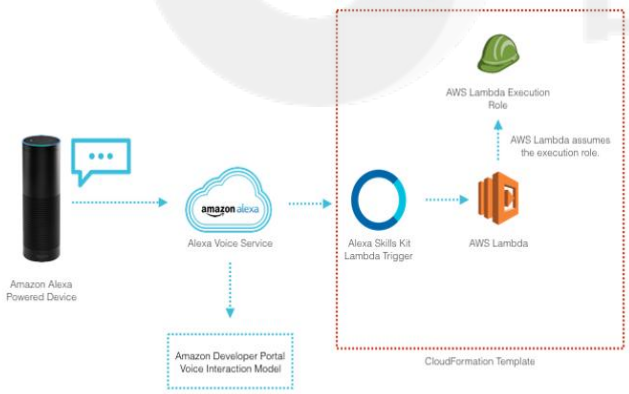
► 比特大陆

比特大陆的计算芯片 BM 1680，集成了深度学习算法硬件加速模块（NPUs），应用于云端计算与推理。BM1680 还提供了 4 个独立的 DDR4 通道，用于高速数据缓存读取，以提高系统的执行速度。其典型功耗只有 25W，在单位能耗推断效率上有一定优势。

应用场景#1：云端推断芯片助力智能语音识别

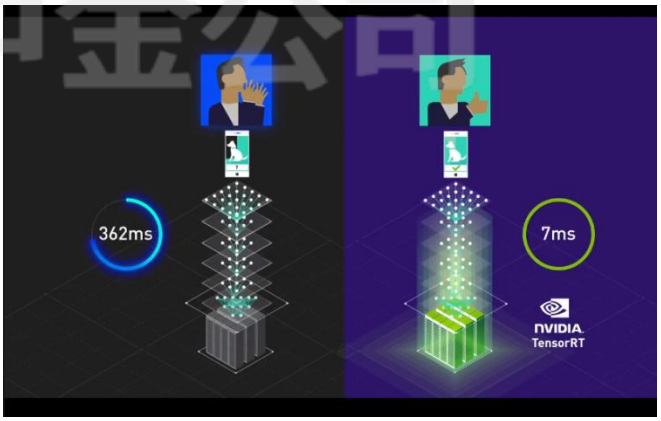
云端推断芯片提升语音识别速度。语音识别是推断芯片的工作场景之一，如 Amazon 的语音助手 Alexa，其“智能”来自于 AWS 云中的推断芯片。Alexa 是预装在亚马逊 Echo 内的个人虚拟助手，可以接收及相应语音命令。通过将语音数据上传到云端，输入推断芯片进行计算，再返回结果至本地来达到与人实现交互的目的。原先云端采用 CPU 进行推断工作，由于算力低，识别中会有 300-400ms 的延迟，影响用户体验。**而现今 AWS 云中采用了 Nvidia 的 P40 推断芯片，结合 Tensor RT 高性能神经网络推理引擎（一个 C++库），可以将延迟缩减到 7ms。**此外，AI 芯片支持深度学习，降低了语音识别错误率。目前，借助云端芯片的良好推断能力，百度语音助手的语音识别准确度已达到 97%之高。

图表 15：智能音箱通过云端推断芯片工作



资料来源：Amazon，中金公司研究部

图表 16：Nvidia 云端推断芯片提升语音识别速度



资料来源：Nvidia，中金公司研究部

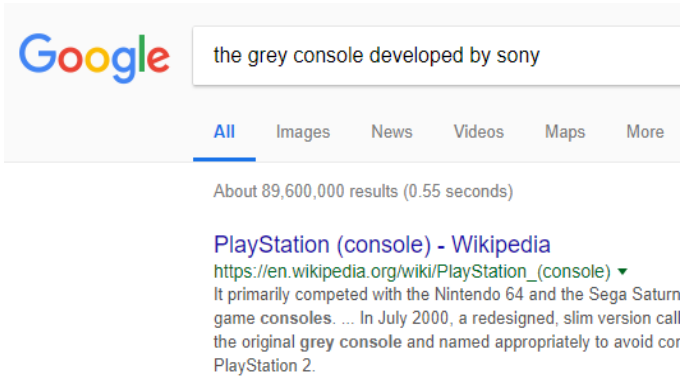
应用场景#2：推断芯片应用于智能搜索

RankBrain 是 Google 众多搜索算法的一部分，它是一套计算机程序，能把知识库中上十亿个页面进行排序，然后找到与特定查询最相关的结果。目前，Google 每天要处理 30 亿条搜索，而其中 15%的词语是 Google 没有见过的。**RankBrain 可以观察到看似无关复杂搜索之间的模式，并理解它们实际上是如何彼此关联的，实现了对输入的语义理解。这种能力离不开 Google 云端推断芯片 TPU 的辅助。**

先前，在没有深度学习情况下，单纯依靠 PageRanking 及 InvertedIndex，Google 也能实现

一定程度的对搜索词条排序的优化，但准确率不够。TPU 利用 RankBrain 中的深度学习模型，在 80%的情况下计算出的置顶词条，均是人们最想要的结果。

图表 17：推断芯片助力深度学习实现语义识别



资料来源：Google，中金公司研究部

图表 18：TPU+RankBrain 在推断正确率上获得提高

	传统服务器	TPU+RankBrain
处理陌生语句	逐词分开搜索	语义理解
推断搜索置顶词条正确率	70%或者更低	80%

资料来源：Google，中金公司研究部

用于智能手机的边缘推断芯片：竞争格局稳定，传统厂商持续受益

手机芯片市场目前包括（1）苹果，三星，华为这类采用芯片+整机垂直商业模式的厂商，以及（2）高通，联发科，展锐等独立芯片供应商和（3）ARM，Synopsys、Cadence 等向芯片企业提供独立 IP 授权的供应商。采用垂直商业模式厂商的芯片不对外发售，只服务于自身品牌的整机，性能针对自身软件做出了特殊优化，靠效率取胜。独立芯片供应商以相对更强的性能指标，来获得剩余厂商的市场份额。

从 2017 年开始，苹果，华为海思，高通，联发科等主要芯片厂商相继发布支持 AI 加速功能的新一代芯片（如下图），AI 芯片逐渐向中端产品渗透。由于手机空间有限，独立的 AI 芯片很难被手机厂采用。在 AI 加速芯片设计能力上有先发优势的企业（如寒武纪）一般通过 IP 授权的方式切入。

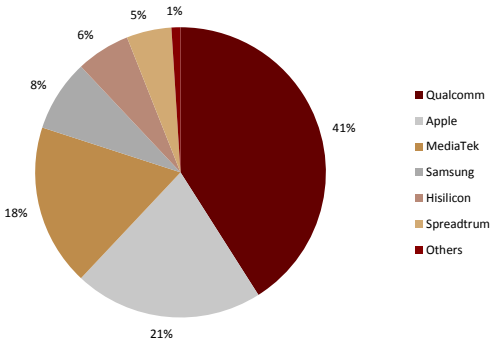
图表 19：手机 AI 芯片对比

SoC 供应商	Apple	Huawei Hisilicon	Samsung	Qualcomm	MediaTek
芯片名称	A11 Bionic	Kirin 970	Exynos 9810	Snapdragon 845	Helios P60
CPU	2x Monsoon+4x Mistral	4x Cortex A73 + 4x Cortex A53	4x M3 (Cortex A75)+ 4x Cortex-A55	4x Kyro 385 Gold+ 4x Kyro 385 Silver	4x Cortex A73+ 4x Cortex A53
GPU	Apple designed 3-core GPU	ARM Mali-G72 MP12	ARM Mali-G72MP12	Adreno 630	ARM Mali-G72MP12
AI处理器	Apple designed 2-core neural engine	NPU	VPU	Hexagon 685 DSP	2 x 140GMACs
内存	LPDDR 4x	LPDDR4	LPDDR4x	LPDDR4x	LPDDR3 LPDDR4x
ISP/摄像头	Apple ISP for faster auto-focus in low-light	Dual 14-bit ISP	Dual-ISP	Dual 14-bit Spectra 280 ISP 1x 32MP or 2x 16MP	1x 32MP or 2x20+16MP
集成通讯模块	NA	Kirin 970 Integrated LTE (Category 18/13)	Custom Cat.18 LTE modem	Snapdragon X20 LTE (Category 18/13)	Category 7/13
制造工艺	TSMC 10nm FinFET	TSMC 10nm FinFET	10nm FinFET	10nm LPP	TSMC 12nm FinFET

资料来源：Intel，Nvidia，Google，中金公司研究部

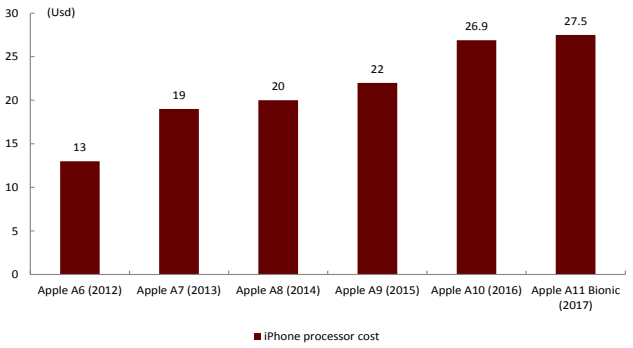
对这些厂商来说，我们认为 AI 化的主要作用是提升芯片附加价值与产品单价。根据 IHS 的数据，随着硬件性能的增强及针对于 AI 的运算结构不断渗透，苹果 A11 芯片的成本已达到 27.5 美元。芯片成本持续上涨有望带动垂直模式厂商整机售价走高，在出货量相同的情况下为现有芯片厂商贡献更多的营业收入。高通、联发科、展锐等独立芯片供应商则会受益于芯片本身 ASP 的提升。

图表 20：智能手机 SoC 市占率分析（2017）



资料来源：CounterPoint，中金公司研究部

图表 21：历代 Apple 手机芯片成本趋势



资料来源：iHS，中金公司研究部

应用场景#1：推断芯片为 AI 拍照技术提供硬件支持

智能手机通过 AI 算法+终端推断芯片，可实现对于现实世界图像的智能识别，并在此基础上进行实时优化：1）从整个场景识别，到特殊优化过程中，推断芯片为算法运行提供硬件支持。2）手机推断芯片中 GPU、NPU 等单元的协同工作，实现了对边缘虚化更准确的处理，使小尺寸感光元件的手机获得“单反”级的景深效果，增加相片的层次感。3）人脸结构的识别也离不开边缘推断芯片，芯片性能的提升直接导致了 AI 美颜、3D 光效等特殊效果变得更加自然。**如果缺少终端芯片的支持，一旦运行高负载的 AI 任务手机就需要呼唤云端。而云端的相应速度不够，导致 AI 摄影的识别率和准确率下降，用户体验将大打折扣。**

应用场景#2：推断芯片助力语音助手处理复杂命令

从“听清”到“听懂”，自然语言理解能力提升与推断芯片硬件的支持分不开：多麦克风方案的普及解决了“听清”的问题，而到“听懂”的跨越中自然语言理解能力是关键。这不仅对云端训练好的模型质量有很高要求，也必须用到推断芯片大量的计算。随着对话式 AI 算法的发展，手机 AI 芯片性能的提升，语音助手在识别语音模式、分辨模糊语音、剔除环境噪声干扰等方面能力得到了优化，可以接受理解更加复杂的语音命令。

图表 22：手机 AI 芯片辅助图片渲染优化

资料来源：OPPO，中金公司研究部

图表 23：手机 AI 芯片辅助 Vivo Jovi 处理复杂命令

资料来源：Vivo，中金公司研究部

用于安防边缘推断芯片：海思、安霸与 Nvidia、Mobileye 形成有力竞争

视频监控行业在过去十几年主要经历了“高清化”、“网络化”的两次换代，而随着 2016 年以来 AI 在视频分析领域的突破，目前视频监控行业正处于第三次重要升级周期——“智能化”的开始阶段。**前端摄像头装备终端推断芯片，可以实时对视频数据进行结构化处理，“云+边缘”的边缘计算解决方案逐渐渗透。**我们预计，应用安防摄像头的推断芯片市场规模，将从 2017 年的 3.3 亿美元，增长至 2022 年的 18 亿美元，CAGR~41%。

应用场景：安防边缘推断芯片实现结构化数据提取，减轻云端压力

即便采用 H.265 编码，目前每日从摄像机传输到云端的数据也在 20G 左右，不仅给存储造成了很大的压力，也增加了数据的传输时间。**边缘推断芯片在安防端的主要应用，基于将视频流在本地转化为结构化数据。**这样既节省云端存储空间，也提升系统工作效率。“视频结构化”，简言之即从视频中结构化提取关键目标，包括车辆、人及其特征等。虽然这种对数据的有效压缩要通过算法实现，但硬件的支持不可或缺。根据海康威视提供

的案例，我们可以看到，由边缘推断芯片支持的结构化分析，可以使原本长达一个月的检索时长缩减到 5 秒内，大幅降低了公安部门的工作强度及难度。

图表 24：视频结构化数据提取实例

资料来源：明景科技，中金公司研究部

图表 25：AI 芯片助力结构化分析实现工作效率提升

	人工分析	海康结构化分析
监控点数量	500	
视频时长	250小时	
检索时长	30天	5秒内
视频中人流量	50万人	
优劣	尝试使用人海战术进行查看 耗时久，易疲惫，可能遗漏关键信息	分析速度快，效率高 节省公安干警办案时间
其他案例剖析	2012年，南京“1.6”周XX抢劫案 监控点：1万多个　　视频：2000T 视频查阅人员：1500多名公安干警　　耗时：1个多月	

资料来源：海康威视，中金公司研究部

传统视频解码芯片厂商积极布局 AI 升级。华为海思、安霸（Ambarella）都在近一年内推出了支持 AI 的安防边缘推断芯片。海思的 HI3559A 配备了双核神经网络加速引擎，并成为第一款支持 8k 视频的芯片；安霸也通过集成 Cvflows 张量处理器到最新的 CV2S 芯片中，以实现对 CNN/DNN 算法的支持。打入视频监控解决方案龙头海康威视，实现前装的 Nvidia，Movidius 同样不甘示弱，Movidius 发布的最新产品 Myriad X 搭载神经计算引擎，在 2W 的功耗下可实现 1TOPS 的算力。Nvidia TX2 是 TX1 的升级产品，算力更强，达到 1.5TFLOPS，存储能力也有提升。

图表 26：安防 AI 芯片对比

供应商名称	华为海思	Movidius	Nvidia	安霸
芯片型号	HI3559A	Myriad X	Jetson TX2	CV2S
发布时间	2017/10	2017/08	2017/03	2018/05
处理器	2x A73 CPU、2x A53 CPU+ Mali G71 GPU+ 4x DSP+2x NNIE神经网络加速引擎	神经计算引擎（NCE）+ 16x 128-bit 流式混合架构向量引擎	2x Denver 2 CPU+ 4x A57 CPU, Pascal 64-bit GPU	4x 64-bit A53 CPU+ 集成Cvflow张量处理器 专门处理CNN/DNN算法
运算能力	0.6 TOPS	1 TOPS	1.5 TFLOPS	NA
功耗	NA	2W	7.5W	NA
特点	提供8K30/4K120 的数字视频录制 支持H.265编码输出 或4K30 RAW视频输出 支持多路4K sensor输入 多路ISP图像处理 内置双目深度检测单元	支持4K60编解码 16路MIPI通道 最多支持连接8个高清摄像机	提供4K60编解码 支持12路CSI通道 最多支持6个摄像机	支持4K60编解码 8路MIPI通道

资料来源：海思半导体，Movidius，Nvidia，安霸，中金公司研究部

我们认为，目前整个安防 AI 芯片市场竞争格局稳定，现有厂商凭借与下游客户长期的合作，有望继续受益于安防智能化的升级，属于新进入者的市场空间有限。安防 AI 芯片下游客户稳定，为海康威视、大华股份等视频监控解决方案提供商。客户与传统视频解码芯片厂商的长期合作具有粘性，同样推出新产品，初创公司的竞争优势弱一些，尤其是在安防 AI 芯片性能差异化很难做到很大的情况下。

用于自动驾驶的边缘推断芯片：一片蓝海，新竞争者有望突围

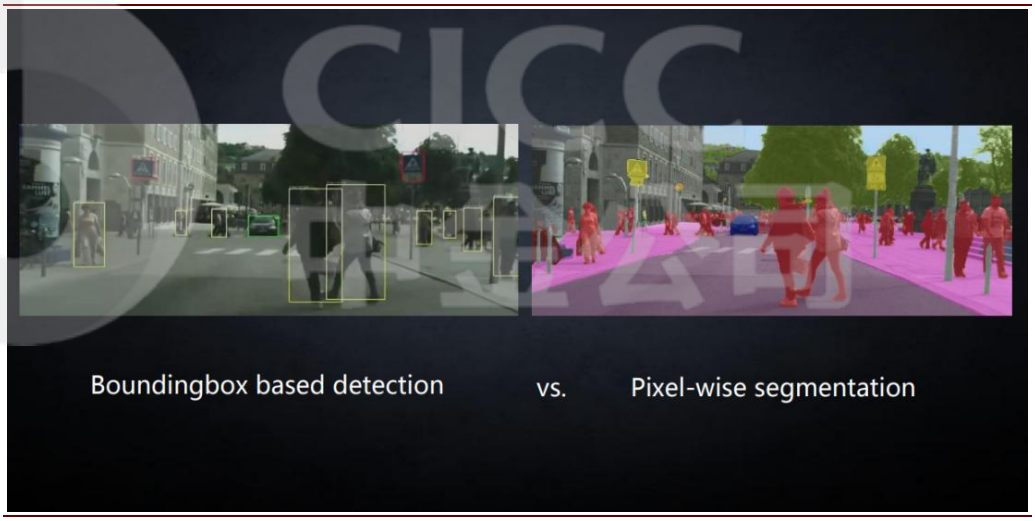
除了智能手机，安防外，自动驾驶汽车也是人工智能的落地场景之一。车用半导体强大需求已经使供给端产能开始吃紧，而用于自动驾驶的推断芯片需求，同样有望在未来 5 年内实现高速增长。我们预计，其市场规模将从 2017 年的 8.5 亿美元，增长至 2022 年的 52 亿美元，CAGR~44%。

若想使车辆实现真正的自动驾驶，要经历在感知-建模-决策三个阶段，每个阶段都离不开终端推断芯片的计算。

应用场景#1：自动驾驶芯片助力环境感知

在车辆感知周围环境的过程中，融合各路传感器的数据并进行分析是一项艰巨的工作，推断芯片在其中起到了关键性作用。我们首先要对各路获得的“图像”数据进行分类，在此基础之上，以包围盒的（bounding box）形式辨别出图像中的目标具体在什么位置。但这并不能满足需求：车辆必须要辨别目标到底是其他车辆，是标志物，是信号灯，还是人等等，因为不同目标的行为方式各异，其位置、状态变化，会影响到车辆最终的决策，因而我们要对图像进行语义分割（segmantation，自动驾驶的核心算法技术）。语义分割的快慢和推断芯片计算能力直接相关，时延大的芯片很显然存在安全隐患，不符合自动驾驶的要求。

图表 27：自动驾驶推断芯片+算法实现视频的像素级语义分割

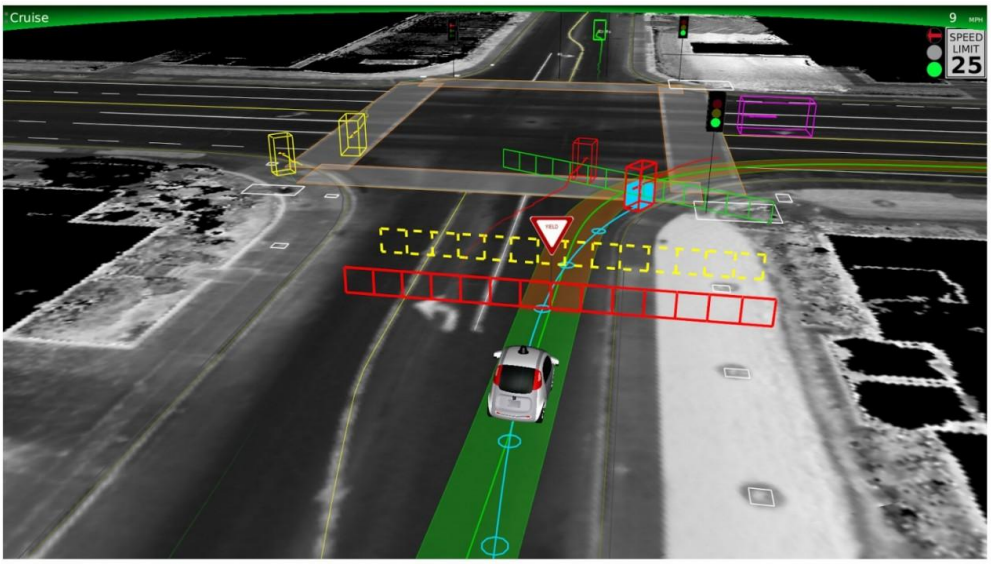


资料来源：地平线机器人，中金公司研究部

应用场景#2：自动驾驶芯片助力避障规划

避障规划是车辆主要探测障碍物，并对障碍物的移动轨迹跟踪（Moving object detection and tracking，即 MODAT）做出下一步可能位置的推算，最终绘制出一幅含有现存、及潜在风险障碍物地图的行为。出于安全的要求，这个风险提示的时延应该被控制在 50ms 级。**随着车速越来越快，无人车可行驶的路况越来越复杂，该数值在未来需要进一步缩短，对算法效率及推断芯片的算力都是极大的挑战。**例如，在复杂的城区路况下，所需算力可能超过 30TOPS。未来 V2X 地图的加入，将基本上确保了无人车的主动安全性，但同样对推断芯片的性能提出了更高的要求。

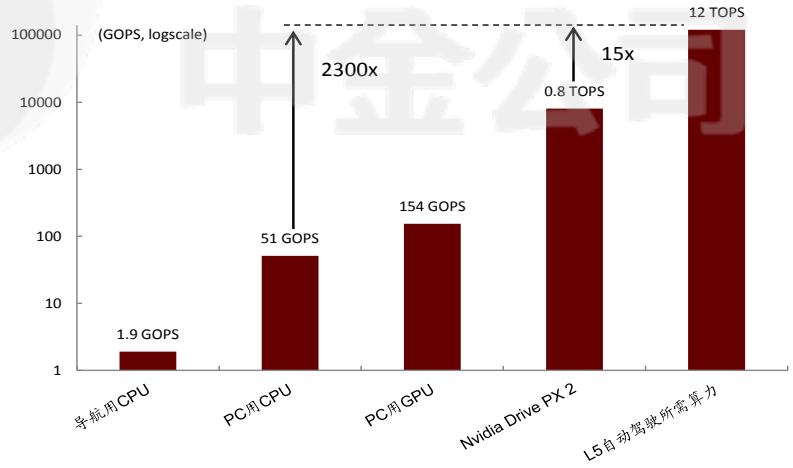
图表 28：自动驾驶推断芯片+算法实现自动驾驶避障规划



资料来源：SlashGear，中金公司研究部

从以上应用场景不难看出，自动驾驶对芯片算力提出了很高的要求，而受限于时延及可靠性，有关自动驾驶的计算不能在云端进行，因此边缘推断芯片升级势在必行。根据丰田公司的统计数据，实现 L5 级完全自动驾驶，至少需要 12TOPS 的推断算力，按现行先进的 Nvidia PX2 自动驾驶平台测算，差不多需要 15 块 PX2 车载计算机，才能满足完全自动驾驶的需求。

图表 29：自动驾驶算力需求加速芯片升级



资料来源：Toyota，中金公司研究部

传统车载半导体厂商积极布局自动驾驶。近些年来，各传统车载半导体供应商纷纷涉猎自动驾驶业务，推出了各自的自动驾驶，或辅助驾驶平台，如 TI 推出了面向于 L1/2 级的平价产品，而 Renesas 和 NXP 步入中高端市场。V3M 与 Bluebox 分别是两家的代表性产品，均满足客户 L3 级自动驾驶需求。目前 NXP 的 Bluebox 2.0 也在测试中。老牌厂商中 Mobileye（被 Intel 收购）在自动驾驶边缘推断芯片上表现最为抢眼，其 EyeQ3 芯片已经被集成于新一代量产 Audi A8 中的 zFAS 平台上，而 A8 也因此成为第一款支持 L3 级自动驾驶的车型。

下一代产品中，Mobileye 和新秀 Nvidia 有望实现领先。Mobileye 更注重算法端，强调软硬件结合带来的效率提升，其开发的最新 EyeQ5 芯片在 10W 的功耗下就能达到 24TOPS 的算力。英伟达作为传统硬件厂商，借助于 GPU 图形处理的优势，也在自动驾驶市场取

得了相应的领先地位，其芯片更注重绝对算力表现。将于今年三季度流片，2019 年三季度量产的“算力怪兽”Pegasus 平台，搭载了两块 Nvidia 下一代的 GPU，将实现 320TOPS 的超强计算能力，完全覆盖 L5 级别应用的需求。

图表 30：自动驾驶平台对比

Vendor	Platform	Processor	Inputs		Features				
			# of Camera	RADAR	Surrounding view	LDW	TSR	FCW	AEB
TI	TDA 2x	2x A15+ 4x M4+ 2x GPU	10	x	x	x		x	
	TDA 3x	2x M4	8	x	x			x	
	TDA 2Eco	1x A5 +4x M4	8		x				
Renesas	R-Car H3	4x A57/A53/Dual lock-step R7+ GPU	8		x	x	x	x	x
	R-Car E2	2x A7/SH-4A core + GPU	8						
	R-Car V2H	2x A15 + GPU	8		x				
	R-Car V3M	2x A53	8	x	x	x	x	x	x
Mobileye/STMicro	EyeQ3	4 MIPS cores + 4VMP cores	4	x		x	x	x	x
	EyeQ4	4 MIPS i-class cores + 1 MIPS m-class core+ 6VMP cores	8	x		x	x	x	x
	EyeQ5	8 MIPS cores + 18VMP cores	20	x		x	x	x	x
Nvidia	PX II	4 Denver + 8x A57 + 2 Parker GPU	12	x	x	x	x	x	x
	Xavier	8x custom ARM+ Volta GPU	16	x	x	x	x	x	x
	Pegasus	16x custom ARM+ 2x next gen GPU	16	x	x	x	x	x	x
NXP	Bluebox	4x A53 + 1x M4 + GPU	20	x	x	x	x	x	x

资料来源：TI，Renesas，Intel，Nvidia，NXP，中金公司研究部

对比其他终端应用场景，自动驾驶不仅计算复杂程度最高，车规级要求也为芯片设立了更高的准入门槛，其硬件升级落地相对缓慢。目前各厂商下一代的自动驾驶平台最早计划于 2019 年量产，现今上市平台中，芯片大多只支持 L2/3 级。之前 Uber 的无人车事故，也对整个行业的发展造成了拖累。

图表 31：下一代自动驾驶 AI 芯片流片及投产时间预估

资料来源：Intel，Nvidia，地平线机器人，中金公司研究部

自动驾驶芯片市场仍处于初期起步阶段。虽然 NXP 等传统半导体厂商深耕于汽车电子多年，获得了一定的客户粘性，但在自动驾驶业务上，整个市场还未形成非常明显的竞争格局。客户也在不断测试芯片厂商的产品，来实现最优选择。根据各公司披露的数据，目前各大芯片厂商与整车厂（OEM）及 Tier 1 厂商都开展了紧密的合作，但客户数量不相上下。从客户的偏好来看，传统大厂愿意自行搭建平台，再采购所需芯片，而新车厂偏向于直接购买自动驾驶平台。介于实现完全自动驾驶非常复杂，目前还在起步阶段，**我们认为初创公司在整个行业的发展中是有机会的，并看好技术领先，能与车厂达成密切合作的初创公司。**

图表 32：各芯片厂商合作方比较

	OEM											Tier 1			
	VW Group	Benz	Audi	BMW	GM	Ford	Nissan	Toyota	Volvo	Nio	Tesla	Bosch	ZF	Valeo	Delphi
NXP	x	x	x	x	x	x			x			x		x	x
Renesas	x	x	x	x			x	x				x		x	x
TI	x		x	x		x									x
Mobileye	x		x	x	x		x			x				x	x
Nvidia	x	x	x			x		x	x	x		x	x		

资料来源：TI，Renesas，Intel，Nvidia，NXP，中金公司研究部

主要中国 AI 芯片公司介绍

中国大陆目前有超 20 家企业投入 AI 芯片的研发中来。除了像华为海思、紫光展锐这种深耕于芯片设计多年的企业之外，也有不少初创公司表现抢眼，如寒武纪、比特大陆等。此外，台湾地区的 GUC（创意电子）是一家 IC 后端设计公司，凭借 20 年的行业经验，和投资方晶圆制造巨头台积电的鼎力支持，在 AI 芯片高速发展的大环境下也有望受益。

图表 33：中国大陆主要 AI 芯片设计公司至少有 20 家

公司名	创始人	成立时间	最新融资轮次	最新融资额	领投方	端 AI 芯片	云 AI 芯片
寒武纪科技	陈天石	2016.3	B轮	1亿美元	国有资本风险投资基金、国投创业、阿里巴巴等	寒武纪1A（应用于华为麒麟970） 寒武纪1H8、1H16和1M	MLU100 MLU200
地平线机器人	余凯	2015.7	A+轮	数千万美元	Intel、嘉实投资、高瓴资本等	旭日（视觉计算） 征程1.0（自动驾驶） 征程2.0（自动驾驶）	
比特大陆	吴忌寒、詹克团	2013.1	A轮	5000万美元	IDG、红杉资本		SOPHON BM1680, BM1682
深鉴科技（Xilinx子公司）	姚颂、汪玉	2016.2	A+轮	4000万美元	蚂蚁金服、三星风投等	嵌入式视觉AI芯片“听涛”	AI芯片“观海”
异构智能（Novumind）	吴韧	2015.11	A轮	1500万美元	洪泰基金、宽带资本、真格基金等	嵌入式视觉AI芯片CI1006	
海康威视	陈宗年	2001.11	n.a.	n.a.	n.a.	2018发改委人工智能创新发展和数字经济试点重大工程拟支持项目“计算机视觉AI芯片研发及产业化项目”	
华为海思	华为子公司	2004.1	n.a.	n.a.	n.a.	2017年9月推出麒麟970手机AI芯片 2018年8月推出麒麟980手机AI芯片 2017年推出Hi3559A V100 安防摄像头芯片	
深思创芯	俞德军	2017.1	天使轮	未透露	清华启迪、成电求实	2018年终推出安防AI芯片	类脑芯片
清华大学	清华大学微电子学研究所所长魏少军教授领导的可重构计算团队，已推出三代低功耗AI芯片Thinker，预计2018年将技术转移至企业进□商业落地						
清华大学	清华精仪系施路平老师组领衔的类脑计算芯片团队						
北京大学	北京大学高能效计算与应用中心主任丛京生带领的基于FPGA的深度学习加速团队						
中星微电子	邓中翰	1999.1	2015年IPO			嵌入式NPU“星光智能1号”	
杭州国芯	黄智杰	2001.1	A轮	1200万美元	未透露	AI语音芯片GX8008	
云天励飞	陈宁	2014.8	A轮	数千万美元	山水从容传媒投资有限公司、松禾资本等	打造AI芯片intellifusion	
西井科技	谭黎敏	2015.5	A轮	未透露	复兴同浩	Deepsouth	
启英泰伦	高君效、何云鹏	2015.11	A轮	3000万人民币	Roobo智能管家等	家电AI芯片CI1006	
耐能（Kneron）	刘峻诚	2015.11	A轮	数千万美元	阿里巴巴创业者基金、红杉资本中国、中科创达、	应用于安防、智能家居的NPU	
海青智盈	黄河、Qi Dong	2017.1	n.a.	300万人民币	美国AI芯片初创公司Gyrfalcon认缴出资成立	AI芯片 Lightspeeur	
云知声	黄伟、康恒	2017.12	n.a.	1000万人民币	云知声认缴出资成立	AI语音芯片Unione、物联网AI芯片雨燕	
华夏芯	李科奕	2014.12	天使轮	未透露	亦庄国投	G60 AI专用IP、多核SoC芯片平台Polaris	
嘉楠耘智	张楠赓	2013.4	挂牌申请新三板	n.a.	n.a.	AI芯片KPU，应用于自动驾驶、语音交互	
眼擎科技	朱继志	2014.3	Pre-A轮	未透露	未透露	“eyemore X42”成像芯片	

资料来源：公司数据，中金公司研究部

我们认为以下企业值得关注：

海思半导体（Hisilicon）

海思半导体成立于 2004 年 10 月，是华为集团的全资子公司。海思的芯片产品覆盖无线网络、固网及数字媒体等多个领域，其 AI 芯片为 Kirin 970 手机 SoC 及安防芯片 Hi3559A V100。Kirin 970 集成 NPU 神经处理单元，是全球第一款手机 AI 芯片，它在处理静态神经网络模型方面有得天独厚的优势。而 Hi3559A V100 是一款性能领先的支持 8k 视频的 AI 芯片。

清华紫光展锐 （Tsinghua UNISOC）

清华紫光集团于 2013 年、2014 年先后完成对展讯及锐迪科微电子的收购，2016 年再将二者合并，成立紫光展锐。紫光展锐是全球第三大手机基带芯片设计公司，是中国领先的 5G 通信芯片企业。Gartner 的数据显示，紫光展锐手机基带芯片 2017 年出货量的全球占比为 11%。除此之外，展锐还拥有手机 AI 芯片业务，推出了采用 8 核 ARM A55 处理器的人工智能 SoC 芯片 SC9863，支持基于深度神经网络的人脸识别技术，AI 处理能力比上一代提升 6 倍。

GUC（台湾创意电子，3443 TT）

公司介绍：GUC 是弹性客制化 IC 领导厂商（The Flexible ASIC Leader™），主要从事 IC 后端设计。后端设计工作以布局布线为起点，以生成可以送交晶圆厂进行流片的 GDS2 文件为终点，需要很多的经验，是芯片实现流片的重要一环。初创公司同时完成前后端设计难度较大。在 AI 芯片设计发展的大环境下，加上大股东台积电的支持，GUC 有望获得大量的后端订单。公司已在台湾证券交易所挂牌上市，股票代号为 3443。

以下为初创公司：

寒武纪科技（Cambricon Technologies）

寒武纪创立于 2016 年 3 月，是中科院孵化的高科技企业，主要投资人为国投创业和阿里巴巴等。公司产品分为终端 AI 芯片及云端 AI 芯片。终端 AI 芯片采用 IP 授权模式，其产品 Cambricon-1A 是全球首个实现商用的深度学习处理器 IP。去年年底公司新发布了第三代机器学习专用 IP Cambricon-1M，采用 7nm 工艺，性能差不多高出 1A 达 10 倍。云端产品上，寒武纪开发了 MLU 100 AI 芯片，支持训练和推断，单位功耗算力表现突出。

比特大陆（Bitmain）

比特大陆成立于 2013 年 10 月，是全球第一大比特币矿机公司，目前占领了全球比特币矿机 60%以上的市场。由于 AI 行业发展迅速及公司发展需要，公司将业务拓展至 AI 领域，并于与 2017 年推出云端 AI 芯片 BM1680，支持训练和推断。目前公司已推出第二代产品 BM1682，相较上一代性能提升 5 倍以上。

地平线机器人（Horizon Robotics）

成立于 2015 年 7 月，地平线是一家注重软硬件结合的 AI 初创公司，由 Intel、嘉实资本、高瓴资本领投。公司主攻安防和自动驾驶两个应用场景，产品为征程 1.0 芯片（支持 L2 自动驾驶）和旭日 1.0（用于安防智能摄像头），具有高性能（实时处理 1080P@30 帧，并对每帧中的 200 个目标进行检测、跟踪、识别）、低功耗（典型功耗在 1.5W）、和低延迟的优势（延迟小于 30 毫秒）。公司二代自动驾驶芯片将于 1Q19 流片，实现语义建模。

云天励飞（Intellifusion）

公司创立于 2014 年 8 月，由山水从容传媒、松禾资本领投，主攻安防 AI 芯片。其自研 IPU 芯片是低功耗的深度学习专用处理器，内含专用图像处理加速引擎，通过级联扩展最多可处理 64 路视频。能耗比突出，超过 2Tops/Watt。

异构智能（NovuMind）

异构智能创立于 2015 年 8 月，由洪泰基金、宽带资本、真格基金和英诺天使投资。2018 年公司展示了其首款云端 AI 芯片 NovuTensor，基于 FPGA 实现，性能已达到目前最先进的桌面服务器 GPU 的一半以上，而耗电量仅有 1/20。公司即将推出的第二款 ASIC 芯片，能耗不超 5W，计算性能达 15 TOPs，将被用于安防和自动驾驶应用中。

龙加智（Dinoplus）

创立于 2017 年 7 月龙加智是专注于云端芯片的 AI 初创公司，由挚信资本和翊翎资本领投。其产品 Dino-TPU 在 75W 功耗下，计算能力超过除最新款 Nvidia Volta 之外的所有 GPU，时延仅为 Volta V100 的 1/10。同时，Dino-TPU 提供市场上独一无二的冗余备份和数据安全保障。公司计划于 2018 年底完成第一款芯片的流片。

法律声明

一般声明

本报告由中国国际金融股份有限公司（已具备中国证监会批复的证券投资咨询业务资格）制作。本报告中的信息均来源于我们认为可靠的已公开资料，但中国国际金融股份有限公司及其关联机构（以下统称“中金公司”）对这些信息的准确性及完整性不作任何保证。本报告中的信息、意见等均仅供投资者参考之用，不构成对买卖任何证券或其他金融工具的出价或征价或提供任何投资决策建议的服务。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐或投资操作性建议。投资者应当对本报告中的信息和意见进行独立评估，自主审慎做出决策并自行承担风险。投资者在依据本报告涉及的内容进行任何决策前，应同时考量各自的投资目的、财务状况和特定需求，并就相关决策咨询专业顾问的意见对依据或者使用本报告所造成的一切后果，中金公司及/或其关联人员均不承担任何责任。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测无需通知即可随时更改。在不同时期，中金公司可能会发出与本报告所载意见、评估及预测不一致的研究报告。

本报告署名分析师可能会不时与中金公司的客户、销售交易人员、其他业务人员或在本报告中针对可能对本报告所涉及的标的证券或其他金融工具的市场价格产生短期影响的催化剂或事件进行交易策略的讨论。这种短期影响的分析可能与分析师已发布的关于相关证券或其他金融工具的目标价、评级、估值、预测等观点相反或不一致，相关的交易策略不同于且也不影响分析师关于其所研究标的证券或其他金融工具的基本面评级或评分。

中金公司的销售人员、交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。中金公司没有将此意见及建议向报告所有接收者进行更新的义务。中金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的意见不一致的投资决策。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现。过往的业绩表现亦不应作为日后回报的预示。我们不承诺也不保证，任何所预示的回报会得以实现。分析中所做的预测可能是基于相应的假设。任何假设的变化可能会显著地影响所预测的回报。

本报告提供给某接收人是基于该接收人被认为有能力独立评估投资风险并就投资决策能行使独立判断。投资的独立判断是指，投资决策是投资者自身基于对潜在投资的目标、需求、机会、风险、市场因素及其他投资考虑而独立做出的。

本报告由受香港证券和期货委员会监管的中国国际金融香港证券有限公司（“中金香港”）于香港提供。香港的投资者若有任何关于中金公司研究报告的问题请直接联系中金香港的销售交易代表。本报告作者所持香港证监会牌照的牌照编号已披露在报告首页的作者姓名旁。

本报告由受新加坡金融管理局监管的中国国际金融（新加坡）有限公司（“中金新加坡”）于新加坡向符合新加坡《证券期货法》定义下的认可投资者及/或机构投资者提供。提供本报告于此类投资者，有关财务顾问将无需根据新加坡之《财务顾问法》第 36 条就任何利益及/或其代表就任何证券利益进行披露。有关本报告之任何查询，在新加坡获得本报告的人员可联系中金新加坡销售交易代表。

本报告由受金融服务监管局监管的中国国际金融（英国）有限公司（“中金英国”）于英国提供。本报告有关的投资和服务仅向符合《2000 年金融服务和市场法 2005 年（金融推介）令》第 19（5）条、 38 条、 47 条以及 49 条规定的人士提供。本报告并未打算提供给零售客户使用。在其他欧洲经济区国家，本报告向被其本国认定为专业投资者（或相当性质）的人士提供。

本报告将依据其他国家或地区的法律法规和监管要求于该国家或地区提供本报告

特别声明

在法律许可的情况下，中金公司可能与本报告中提及公司正在建立或争取建立业务关系或服务关系。因此，投资者应当考虑到中金公司及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。

与本报告所含具体公司相关的披露信息请访问 http://research.cicc.com/disclosure_cn，亦可参见近期已发布的相关个股报告。

与本报告所含具体公司相关的披露信息请访 https://research.cicc.com/footer/disclosures，亦可参见近期已发布的关于该等公司的具体研究报告。

研究报告评级分布可从https://research.cicc.com/footer/disclosures 获悉。

个股评级标准：分析员估测未来6~12 个月绝对收益在20%以上的个股为“推荐”、在-10%~20%之间的为“中性”、在-10%以下的为“回避”。星号代表首次覆盖或再次覆盖。

行业评级标准：“超配”，估测未来6~12 个月某行业会跑赢大盘10%以上；“标配”，估测未来6~12 个月某行业表现与大盘的关系在-10%与10%之间；“低配”，估测未来 6~12 个月某行业会跑输大盘 10%以上。

V160908
编辑：江薇、樊荣

北京

中国国际金融股份有限公司
北京市建国门外大街 1 号
国贸写字楼 2 座 28 层
邮编：100004
电话：(86-10) 6505-1166
传真：(86-10) 6505-1156

上海

中国国际金融股份有限公司上海分公司
上海市浦东新区陆家嘴环路 1233 号
汇亚大厦 32 层
邮编：200120
电话：(86-21) 5879-6226
传真：(86-21) 5888-8976

香港

中国国际金融（香港）有限公司
香港中环港景街 1 号
国际金融中心第一期 29 楼
电话：(852) 2872-2000
传真：(852) 2872-2100

深圳

中国国际金融股份有限公司深圳分公司
深圳市福田区益田路 5033 号
平安金融中心 72 层
邮编：518000
电话：(86-755) 8319-5000
传真：(86-755) 8319-9229

Singapore

China International Capital Corporation (Singapore) Pte. Limited
#39-04, 6 Battery Road
Singapore 049909
Tel: (65) 6572-1999
Fax: (65) 6327-1278

United Kingdom

China International Capital Corporation (UK) Limited
Level 25, 125 Old Broad Street
London EC2N 1AR, United Kingdom
Tel: (44-20) 7367-5718
Fax: (44-20) 7367-5719

北京建国门外大街证券营业部
北京市建国门外大街甲 6 号
SK 大厦 1 层
邮编：100022
电话：(86-10) 8567-9238
传真：(86-10) 8567-9235

上海黄浦区湖滨路证券营业部
上海市黄浦区湖滨路 168 号
企业天地商业中心 3 号楼 18 楼 02-07 室
邮编：200021
电话：（86-21) 56386-1195、6386-1196
传真：（86-21) 6386-1180

南京汉中路证券营业部
南京市鼓楼区汉中路 2 号
亚太商务楼 30 层 C 区
邮编：210005
电话：(86-25) 8316-8988
传真：(86-25) 8316-8397

厦门莲岳路证券营业部
厦门市思明区莲岳路 1 号
磐基中心商务楼 4 层
邮编：361012
电话：(86-592) 515-7000
传真：(86-592) 511-5527

重庆洪湖西路证券营业部
重庆市北部新区洪湖西路 9 号
欧瑞蓝爵商务中心 10 层及欧瑞
蓝爵公馆 1 层
邮编：401120
电话：(86-23) 6307-7088
传真：(86-23) 6739-6636

佛山季华五路证券营业部
佛山市禅城区季华五路 2 号
卓远商务大厦一座 12 层
邮编：528000
电话：(86-757) 8290-3588
传真：(86-757) 8303-6299

宁波扬帆路证券营业部
宁波市高新区扬帆路 999 弄 5 号
11 层
邮编：315103
电话：(86-0574) 8907-7288
传真：(86-0574) 8907-7328

北京科学院南路证券营业部
北京市海淀区科学院南路 2 号
融科资讯中心 B 座 13 层 1311 单元
邮编：100190
电话：(86-10) 8286-1086
传真：(86-10) 8286-1106

深圳福华一路证券营业部
深圳市福田区福华一路 6 号
免税商务大厦裙楼 201
邮编：518048
电话：(86-755) 8832-2388
传真：(86-755) 8254-8243

广州天河路证券营业部
广州市天河区天河路 208 号
粤海天河城大厦 40 层
邮编：510620
电话：(86-20) 8396-3968
传真：(86-20) 8516-8198

武汉中南路证券营业部
武汉市武昌区中南路 99 号
保利广场写字楼 43 层 4301-B
邮编：430070
电话：(86-27) 8334-3099
传真：(86-27) 8359-0535

天津南京路证券营业部
天津市和平区南京路 219 号
天津环贸商务中心（天津中心）10 层
邮编：300051
电话：(86-22) 2317-6188
传真：(86-22) 2321-5079

云浮新兴东堤北路证券营业部
云浮市新兴县新城镇东堤北路温氏科技园服务
楼 C1 幢二楼
邮编：527499
电话：(86-766) 2985-088
传真：(86-766) 2985-018

福州五四路证券营业部
福州市鼓楼区五四路 128-1 号恒力城办公楼
38 层 02-03 室
邮编：350001
电话：(86-591) 8625 3088
传真：(86-591) 8625 3050

上海浦东新区世纪大道证券营业部
上海市浦东新区世纪大道 8 号
上海国金中心办公楼二期 46 层 4609-14 室
邮编：200120
电话：(86-21) 2057-9499
传真：(86-21) 2057-9488

杭州教工路证券营业部
杭州市教工路 18 号
世贸丽晶城欧美中心 1 层
邮编：310012
电话：(86-571) 8849-8000
传真：(86-571) 8735-7743

成都滨江东路证券营业部
成都市锦江区滨江东路 9 号
香格里拉办公楼 1 层、16 层
邮编：610021
电话：(86-28) 8612-8188
传真：(86-28) 8444-7010

青岛香港中路证券营业部
青岛市市南区香港中路 9 号
香格里拉写字楼中心 11 层
邮编：266071
电话：(86-532) 6670-6789
传真：(86-532) 6887-7018

大连港兴路证券营业部
大连市中山区港兴路 6 号
万达中心 16 层
邮编：116001
电话：(86-411) 8237-2388
传真：(86-411) 8814-2933

长沙车站北路证券营业部
长沙市芙蓉区车站北路 459 号
证券大厦附楼三楼
邮编：410001
电话：(86-731) 8878-7088
传真：(86-731) 8446-2455

西安雁塔证券营业部
西安市雁塔区二环南路西段 64 号
凯德广场西塔 21 层 02/03 号
邮编：710065
电话：（+86-29）8648 6888
传真：（+86-29）8648 6868

CICC
中金公司