AI芯片洗牌 演绎哪些新趋势

本报记者 张心怡

无芯片不AI，芯片是支撑人工智能的基础。2019年，云端AI芯片迎来亚马逊、高通、阿里巴巴、Facebook等新玩家，软硬一体化趋势加强；终端芯片功耗比竞争加剧，语音芯片持续火热；边缘AI芯片势头初现。2020年，AI芯片将逐渐进入洗牌期，机遇与挑战并存。

边缘AI芯片进入抢滩战

围绕边缘AI芯片的抢滩布局已经开始，头部厂商正在打造云、边、端一体化的计算格局。

AI正在从云端向边缘端扩展，边缘计算被视为人工智能的下一个战场。寒武纪副总裁刘道福表示，在边缘计算种类中，边缘往往和各类传感器相连，而传感器的数据往往是非结构化的，很难直接用于控制和决策，因此需要边缘人工智能计算将非结构化的数据结构化，从而用于控制和决策。

2019年，围绕边缘AI芯片的抢滩布局已经开始。一方面，英伟达、寒武纪、百度等已经在云、端有所积累的厂商，希望以边缘芯片完善云、边、端生态，打造一体化的计算格局。

英伟达发布了面向嵌入式物联网的边缘计算设备Jetson Nano，适用于入门级网络硬盘录像机、家用机器人以及具备全面分析功能的智能网关等应用，之后英伟达又发布了边缘AI超级计算机Jetson Xavier NX，能够在功耗10W的模式下提供最高14TOPS，在功耗15W模式下提供最高21TOPS的性能。

寒武纪发布了用于深度学习的SoC边缘加速芯片思元220，采用台积电16nm工艺，最大算力32TOPS（INT4），功耗控制在10W，支持Tensorflow、Caffe、mxnet以及pytorch等主流编程框架。

百度联合三大运营商、中兴、爱立信、英特尔等企业，发起百度AI边缘计算行动计划，旨在利用AI推理、函数计算、大数据处理和产业模型训练推动AI场景在边缘计算的算力部署和平台支持。

另一方面，自动驾驶等专用边缘AI芯片势头渐显。地平线宣布量产国内首款车规级AI芯片“征程二代”，采用台积电28nm工艺，可提供超过4TOPS的等效算力，典型功耗控制在2瓦，延迟少于100毫秒，多任务模式下可同时运行超过60个分类任务，每秒识别目标数超过2000个，以应对车联网对强实时响应的需求。

多个新玩家入局云端

软件是智能操作的核心，随着异构计算逐渐导入AI芯片，软硬件协同成为云端AI的重要趋势。

云端仍然是AI芯片的主要战场。2019年，云端芯片迎来多个新玩家，算力大战持续升级。

高通推出了面向数据中心推理计算的云端AI芯片Cloud AI 100，峰值性能超过350TOPS，相比其他商用方案每瓦性能提升10倍。

云服务领跑者亚马逊推出了机器学习推理芯片AWS Inferentia，最高算力为128TOPS，在AI推理实例inf1可搭载16个Inferentia芯片，提供最高2000TOPS算力。

阿里巴巴推出号称“全球最高性能AI推理芯片”含光800，采用自研芯片架构和达摩院算法，在Resnet50基准测试中获得单芯片性能第一。

腾讯投资的燧原科技发布了面向云端数据中心的AI加速卡云燧T10，单卡单精度算力达到20TOPS，支持单精度FP32和半精度BF16的混合精度计算，并为大中小型数据中心提供了单节点、单机柜、集群三种模式，在集群模式下可通过片间互联实现1024节点集群。

芯片是AI的载体，而软件是完成智能操作的核心。随着异构计算逐渐导入AI芯片，软硬件协同成为云端AI的重要趋势。英特尔推出了面向异构计算的统一软件平台One API，以隐藏硬件复杂性，根据系统和硬件自动适配功耗最低、性能最佳的加速方式，简化并优化编程过程。赛灵思也推出了软件平台Vitis AI，向用户开放易于访问的软件接口，可根据软件或算法自动适配赛灵思硬件架构。

功耗比仍是终端侧重点

AI正在渗入手机和语音芯片，反映了AI芯片在特定领域的专业化、定制化趋势。

在终端侧，功耗比仍然是角逐焦点。尤其在手机等对于续航能力锱铢必较的终端，主力厂商推出的AI引擎都对低功耗有所强调。

麒麟990 5G的NPU采用双大核+微核的方式，大核负责性能，微核拥有超低功耗。据介绍，微核在人脸检测的应用场景下，能耗是大核工作的1/24。高通发布的骁龙865集成了传感器中枢，让终端能够以极低功耗感知周围情境。三星提出通过较低功耗的NPU实现终端设备上的AI处理，实现在设备端直接执行更复杂的任务。

除了手机，终端侧的另一个当红炸子鸡是AI语音芯片。科大讯飞、阿里巴巴、探境科技、清微智能等都发布了针对智能家居的AI语音芯片，反映了AI芯片在特定领域的专业化、定制化趋势。阿里达摩院公布了首款专用于语音合成算法的AI FPGA芯片技术Ouroboros，使用了端上定制硬件加速技术，降低对云端网络的依赖，支持实时语音合成和AI语音识别，有望率先在天猫精灵搭载。

2020机遇挑战并存

AI芯片将持续火热，企业扎堆进入。但是2020年前后，将出现一批出局者，行业洗牌开始。

2019—2021年，中国AI芯片市场规模仍将保持50%以上的增长速度，到2021年，市场规模将达到305.7亿元。赛迪智库预测，2019—2021年，云端训练芯片增速放缓，云端推理芯片、终端推理芯片市场增长速度将持续上升。预计2021年，中国云端训练芯片市场规模将达到139.3亿元，云端推理芯片市场规模将达到82.2亿元，终端推理芯片市场规模将达到84.1亿元。

集邦咨询分析师姚嘉洋向记者指出，2019年，AI芯片大致走出了一条较为清晰的道路，端、边、云的芯片规格相对明确。2020年，各大芯片厂会延续2019年的产品发展路径，持续深化芯片的性价比及功耗比表现。从训练端来看，值得关注的是HBM（高频宽存储器）的整合与相关的封装技术良率，这会牵动芯片厂商与存储器及封测厂商之间合作关系的变化。推理端的决战点在INT8领域，重点在于如何进一步提升芯片本身的性能及功耗表现。

5G、VR/AR等新技术，也将为AI芯片，尤其是边缘侧的AI芯片提供更多发挥空间。Arm ML事业群商业与营销副总裁Dennis Laudick曾向记者表示，5G通信技术改变了数据处理方式，让边缘AI的工作负载有了处理需求。可以说，5G带来了网络边缘的更多创新。姚嘉洋也表示，AI在5G核心网络存在机会，由于5G带来了更多元的频谱组合，AI可以辅助核心网络更有效地调度网络资源，将频宽资源的利用达到极大化。同时，5G也涵盖车联网，AI在自动驾驶方面将大有机会。在VR/AR端，AI也在导入，主要聚焦在人眼追踪或是场景识别等应用，有望改善VR/AR的流畅度与实时性表现。

清华大学微电子所所长魏少军表示，从产业发展规律来看，在2019-2020年，AI芯片将持续火热，企业扎堆进入；但是2020年前后，将出现一批出局者，行业洗牌开始。由于目前AI算法还在不断演进汇总的过程中，最终的成功与否将取决于各家技术路径的选择和产品落地的速度。

痛点尚待攻克

AI芯片短期内以异构计算为主，中期要发展自重构自学习，长期则朝向通用计算芯片发展。

近两年，AI在语音识别、图像识别等应用领域取得突破，但要从单点突破走向全面开花，需要AI领域诞生如同CPU一样的通用AI计算芯片。清华大学微纳电子系副教授尹首一等专家指出，AI芯片短期内以异构计算为主，中期要发展自重构、自学习、自适应，长期则朝向通用计算芯片发展。

具体来说，AI要从应用适应硬件走向硬件适应应用，就要求AI芯片具备可编程性及动态可变的计算架构，来应对层出不穷的新算法和新应用。魏少军表示，AI芯片一要适应算法的演进，二要有适应所有应用的架构，这就要求架构具备高效的转化能力。在成本敏感的消费电子领域，还需关注AI芯片的计算效能，达到低功耗、小体积、开发简易，这些都需要架构上的创新。

全球AI芯片产业仍处于产业化早期阶段，国产处理器厂商与国际厂商在人工智能这一全新赛场上处在同一起跑线。耐能创始人兼CEO刘峻诚表示，中国拥有庞大的智能手机、智能家居、智能安防市场，对中国的AI公司而言，不仅在服务国内客户时具有本土化优势，还可借助这些客户的生产制造优势进军海外市场，实现“立足中国，放眼全球”的商业布局。

厦门联芯获ISO15408-EAL6安全认证 成为国内首家获认证的代工企业

本报记者 陈炳欣

联芯集成电路制造（厦门）有限公司（简称“联芯”）近日宣布，其厂区取得德国联邦信息安全局（Germany Federal Office for Information Securiy，BSI）ISO 15408安全认证EAL6级，成为中国本土首家获此认证的芯片代工企业。这意味着厦门联芯可以提供最高安全等级的半导体制造代工服务，将有效扩展其在金融IC、电子证件等市场的发展。

据悉，ISO 15408是针对信息安全相关产品或系统所制定的安全评估准则，由德国BSI主导，已成为国际标准认证，也是全球最严谨的安全系统评估准则。ISO 15408安全验证等级由低至高共分为EAL1至EAL7，目前EAL6为半导体制造业可认证的最高等级。联芯获得EAL6等级认证殊荣，代表联芯所提供的生产过程安全防护，已符合大多数安全产品的需求，涵盖电子证件、电子银行、电子支付等各类应用。

ISO 15408分为产品认证（Product Certification）与现场验证（Site Certification）两个类别。产品认证系针对个别产品，每一安全产品都需通过认证，方可取得各自产品证书。现场验证则是针对作业厂址，属于作业厂址所属公司所拥有的认证体系。联芯通过现场验证表示其厂区已达到一定的安全标准，在认证效期内为客户生产的所有安全产品，可无须重复认证。

随着人们对信息安全重视程度不断提高，电子证件、电子银行、电子支付对安全芯片的需求也越来越多。以银行金融IC卡为例，为了促进中国银行卡的产业升级和可持续发展，中国金融系统全面推进金融IC卡的应用。2015年1月1日起，全国各银行全面发行银联金融IC卡，磁条卡将逐步退出历史舞台。根据中国人民银行部署的金融IC卡应用工作时间表，2015年年底，110个金融IC卡公共服务领域应用城市POS终端非接受理比例同比至少增加20个百分点；自2016年1月1日起，发卡银行、银行卡清算机构等开展的移动金融服务以基于金融IC卡芯片的有卡交易方式为主。在人民银行的政策指导下，金融IC卡市场的发展前景一片光明。

此外，交通卡、ETC智能卡市场也在不断扩大。2019年我国进入5G商用元年。随着5G通信时代的来临，通信智能卡、物联网领域eSIM产品市场将有望迎来新的发展机遇。根据中国半导体行业协会集成电路设计分会理事长魏少军在ICCAD上发布的数据，中国智能卡企业从2018年的71家增加到102家，销售额提升了24.6%，达到172.1亿元。

联芯首席执行官暨副董事长许志清表示：“ISO15408认证是取得客户信赖的基础，也是公司拓展业务的一个重要通行证。很荣幸联芯成为首家获得ISO15408 EAL6等级殊荣的本土芯片代工企业，证明了联芯在生产制造安全产品时的严谨安全防护已经获得了国际级的肯定与认可。这将协助我们的客户节省产品认证时间，加速其产品上市时程，使联芯成为广大客户的上佳合作伙伴。”

联发科再推5G芯片

本报讯 继去年11月发布天玑1000 5G旗舰芯片，联发科近日再度发布5G芯片天玑800。该芯片采用7nm制程，集成5G基带，面向中高端手机市场。

据介绍，天玑800系列高度集成了联发科的5G调制解调器，支持5G Sub-6GHz频段、5G SA/NSA双模组网，支持2G-5G四代蜂窝连接、动态频谱共享技术以及VoNR语音服务，并号称能效更高。相比外挂解决方案，天玑800系列可显著降低功耗。

天玑800系列采用了4个主频为2GHz的“大核”Cortex-A76，以及4个主频为2GHz的高能效Cortex-A55核心。同时，天玑800集成了4个G77 GPU图形核心，采用旗舰级图像信号处理器（ISP），最多可支持四个摄像头，支持6400万像素传感器和各类多摄像头组合。

天玑800系列还采用了独特的4核架构APU3.0，由3种不同类型的核心组成，可提供高达2.4 TOPS的AI性能。

联发科无线通信事业部总经理李宗霖表示，联发科已推出旗舰级的5G智能手机芯片天玑1000系列，将通过天玑800系列将5G带入中端和大众市场。

据悉，首批搭载天玑800系列5G芯片的终端将于2020年上半年上市。（文 编）

英特尔发布首款消费级独立显卡

本报讯 英特尔首次公开演示了代号为DG1的消费级独立显卡，这也是英特尔时隔多年再次推出此类产品。据介绍，DG1基于Xe架构，即将发布的10nm Tiger Lake芯片也将在集成显卡使用同样的图形架构。

据悉，Intel DG1是英特尔第一款针对消费级平台的独立显卡产品，集成了媒体与显示引擎，专为功耗优化平台设计，并针对游戏和内容创作进行优化。

与上一代芯片相比，Tiger Lake的集成Xe显卡的图形性能将提高一倍，这意味着DG1 GPU可能会给英特尔的图形性能带来更大的提升。据报道，英特尔希望利用其CPU/GPU同步的解决方案，使集成和独立显卡协同工作，让图像处理能力翻倍。

欧亚经济委员会数据库数据显示，Xe DG1独显搭载了96个执行单位，有望搭载768个渲染单元，符合入门级独显的定位。（文 编）

社长：卢山 社址：北京市海淀区紫竹院路66号赛迪大厦18层 邮编：100048 每周二、五出版 周二8版 周五8版 零售4.50元 全年定价420元（包含手机客户端） 广告部：010-88558848/8808 发行部：010-88558777 京昌工商广登字第20170004号 经济日报印刷厂印刷