CAICT 中国信通院

先进计算发展研究报告

（2018 年）

CAICT 中国信通院

中国信息通信研究院

2018年12月

版权声明

本研究报告版权属于中国信息通信研究院，并受法律保护。转载、摘编或利用其它方式使用本研究报告文字或者观点的，应注明“来源：中国信息通信研究院”。违反上述声明者，本院将追究其相关法律责任。

CAICT 中国信通院

前　言

作为信息技术领域的基础和核心，近半个世纪以来，不断涌现的计算技术浪潮推动着信息技术的持续发展和普及，对国民经济、社会发展发挥着基础性、渗透性、引领性的作用，是科技、经济和社会创新发展的重要推动力。

现代计算技术起源于 1946 年埃尼阿克(ENIAC)的诞生，迄今为止已有超过七十年的发展历程，在半导体工艺器件、CPU 芯片以及分布式、集群、异构等系统技术的推动之下，历经大型机/小型机、PC/服务器、集群/分布式、小型化/低功耗等发展阶段，成为 ICT 产业升级最快、创新最活跃的领域之一。

供给和需求的不匹配推动计算技术产业进入新一轮发展周期，人工智能、自动驾驶、物联网、VR/AR 等创新应用取代基础软硬件成为创新新动能，面向不同应用计算需求的优化和加速将成为近期先进计算技术升级的主要思路。当前，技术创新模式和产业生态体系重构在即，开源开放的影响力日益凸显，多元化的生态发展趋势也为后进入者带来更多的发展机遇。

目 录

CAICT 中国信通院

一、 先进计算的内涵和体系

（一） 计算技术产业发展历程

计算技术与人类文明同期起步，历经手动、机械、电动及电子四大阶段。手动计算阶段最早可溯源至远古时代，人类早期通过手指或石子、木棍、结绳等工具实现计数和简单计算，后期随着数学理论的发展逐渐衍生出算筹、算盘、计算尺等计算工具，此时计算的实现以人自身的逻辑计算为主，工具只是辅助实现手段。十七世纪初，伴随人类机械制造能力的不断进步，可用于实现计算的机械装置也日益复杂，包括计算钟、计算器、差分机、分析机等多种基于齿轮传送等机械原理实现的计算机层出不穷，相比较手动阶段而言，极大提升了计算能力。十九世纪末，电机工程学的进步和电的发现及使用给人类社会带来深刻影响的同时，也推动了机械式计算机的又一次进步，电力不仅成为计算设备的动力来源，带动包括制表机、祖思机等一系列复杂计算装置的发展，也推动了后续二进制数字计算机的快速实现。二十世纪四十年代后，诞生于美国的埃尼阿克(ENIAC)标志着计算电子化时代的开启，自此后七十多年里，计算技术的性价比保持指数级增速，成为科技创新、社会进步和经济增长的重要驱动力，据研究机构表明，数字化程度每提高 10%，人均 GDP 增长 0.5-0.6%。

自动化电子化推动计算技术的规模普惠和计算产业的快速繁荣。计算技术是伴随人类实践的需求演进而逐步发展的，按照物理实现手段的不同，可分为电子管、晶体管和集成电路三个阶段。因二战中弹道计算等需求的激增，各国均加大对计算技术的研发投入力度，科学

与军事应用需求推动计算设备的快速发展，先是电子管因超过千倍的开关速度提升取代继电器成为计算机的核心运算部件，而后晶体管计算机因在开关速度、省电和使用寿命的更优表现取代电子管计算机成为上世纪六十年代后期的计算主体设备。此后，各国因二战后发展经济的需求，将原用于军事领域的计算技术应用到国民经济重要行业，使得其在银行、保险、股票等文书类工作量较大的行业中快速普及，开启计算技术由军用科研向民用领域渗透的前奏，在公共服务领域的深入应用最终推动计算技术的快速繁荣和大量普及。

目前计算技术以冯诺依曼架构为基础，围绕数据处理、数据存储、数据交互三大能力要素不断演进升级。1945 年冯·诺依曼正式提出计算机体系架构，后被广泛称之为“冯诺依曼体系架构[1]”，主要内容包括三方面：一是采用二进制进行计算；二是基于存储程序控制理念，计算机按照预先编制好的程序顺序执行完成计算过程；三是计算设备包括运算器、控制器、存储器、输入装置和输出装置五大组成部件。从每秒可进行数千次计算的埃尼阿克(ENIAC)起，到至今每秒已达到数亿亿次运算的中国“神威•太湖之光”超级计算机，计算技术在遵循冯诺依曼体系结构的前提下，围绕数据处理、数据存储和数据交互展开了快速创新迭代。数据处理方面，集成了控制器和运算器功能的中央处理器 CPU 成为计算系统的核心，并逐渐引入图形处理器 GPU、数字信号处理器 DSP、现场可编程门阵列 FPGA 等多样化运算器单元。数据存储方面，随着汞延迟线、穿孔卡片、磁带、动态随机存取内存

[1]首次发表于冯•诺依曼与戈德斯坦、勃克斯等联名发表《First Draft of a Report on the EDVAC》，即计算机史上著名的“101 页报告”中。

DRAM、软盘、硬盘、闪存等存储介质的存储密度、读写效率不断发展的同时，整体存储架构也在快速变化，历经总线架构、交换式架构、矩阵直连架构、分布式架构、全共享交换式架构等多种，推动数据存储的高性能、高可靠和灵活扩展升级。数据交互方面，包括单计算设备内部的总线技术，以及多计算设备间数据互通的以太网技术等均围绕高速率、高带宽、低延时等方面升级数据交换能力，提升整体计算系统的效能表现。

（二） 对先进计算的理解

先进计算并非特指某项具体的计算技术，而是面向未来的多种计算技术的统称。现阶段基于不同层面、不同角度、不同应用场景的计算技术创新层出不穷，各种计算技术、产品及概念不断涌现，从与技术创新相关的专业领域角度来看，先进计算技术创新将是涵盖原理、材料、工艺、器件、系统、算法、网络架构、应用等在内的系统工程，在不同阶段将具有不同的发展特征和发展重点。**短期来看，基于硅基冯诺依曼架构的现代计算技术仍然构成先进计算的主体，面向不同应用需求的系统优化成为技术创新重点方向，**器件及芯片、系统技术和应用技术等将同步发展。**长期而言，因硅基集成电路的物理极限和冯诺依曼架构的固有瓶颈，量子/类脑等非冯诺依曼架构计算技术的突破和产业化将是支撑先进计算未来持续快速升级的重要动力。**

现代计算技术演进至今，已形成相对清晰的技术分层体系。主要包括基础理论、器件技术、部件技术、系统技术和应用技术等五大部分。其中，**基础理论层**是指奠定现代计算技术的理论基础，阿兰·图

灵提出可计算理论和计算机通用逻辑模型——“图灵机”，到目前为止依然是评判可计算性的唯一模型；香农提出可运用布尔理论实现数学问题、逻辑问题和物理实现间的映射，是采用二进制实现计算技术的理论指导；冯·诺依曼提出计算机的构成要素及运作机制，成为实现现代计算机的核心架构。**器件技术层**是指构成计算设备和计算系统所需的电子器件技术，目前主要指与超大规模集成电路实现相关的设计、制造及封测技术。**部件技术层**包括构成计算设备和计算系统的芯片、模块等，主要可分为计算部件、存储部件和通信部件等三大单元，计算部件指 CPU、GPU 和 FPGA 等数据处理硬件，存储部件指内存、外存等数据存储硬件，通信部件是计算部件和存储部件间实现数据交互的硬件。**系统技术层**是指面向不同应用场景需求构建多样化计算系统所需的系统架构、互联架构、存储架构等硬件技术和资源管理、任务调度等软件技术。现阶段对计算系统的分类并无统一定义，根据任务调度模式的不同可分为集中式计算和分布式计算等，根据计算资源种类的不同可分为异构计算和可重构计算等，根据计算所需数据存储位置的不同可分为内存计算和存算一体化等，面向不同应用需求的计算系统技术不仅存在较大差异，且存在融合发展的趋势。**应用技术层**是指多类应用所需的通用功能性技术，目前主要包括数据库、图形图像处理、数字多媒体、安全防护等。

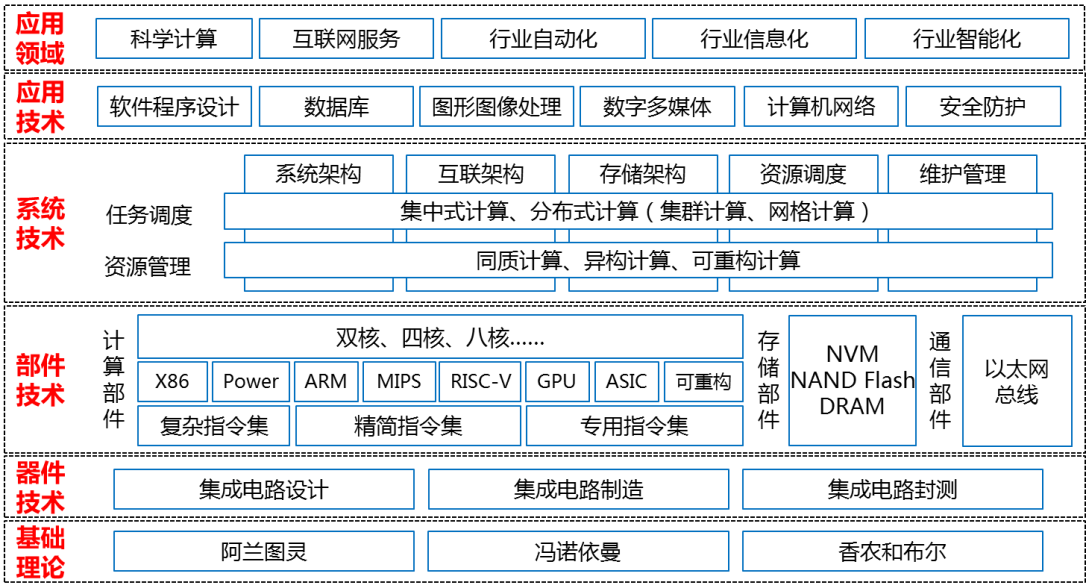


图1 基于冯诺依曼的现代计算技术体系

量子及类脑等非冯诺依曼架构计算技术体系仍未固定，并将因基础原理、物理材料等多方面的巨大差异而较难形成统一体系。通过颠覆冯诺依曼架构，开拓计算技术发展的新路径始终是业界的探索方向，现阶段量子和类脑是发展的热点。整体而言，非冯诺依曼架构将与冯诺依曼架构存在巨大差异，现阶段围绕量子和类脑两大热点的探索实现也非常多样化，在基础理论、物理实现、核心硬件、算法软件等诸多环节均未形成统一定论，非冯诺依曼计算技术整体发展仍处于较为初期的阶段，在可预见的很长时间内，其仍将基于与现代计算体系融合发展的理念进行推进。

综上所述，先进计算涉及面广、技术要素庞杂，2018 年本研究报告将重点围绕未来 3-5 年先进计算领域的核心技术展开分析，探讨技术创新的总体态势和发展重点，并对量子计算、类脑计算等目前业界较为关注的颠覆性技术进行研判。

二、 现阶段先进计算总体发展态势

（一） 供需不匹配是目前计算技术产业面临的主要问题

1. 固有计算技术升级模式遭遇天花板

二十一世纪以来，计算技术升级速度逐渐放缓，起因于芯片主频提升、多核数目堆叠、工艺尺寸微缩等固有升级路径因遭遇瓶颈而渐次失效，主要体现在以下四方面：

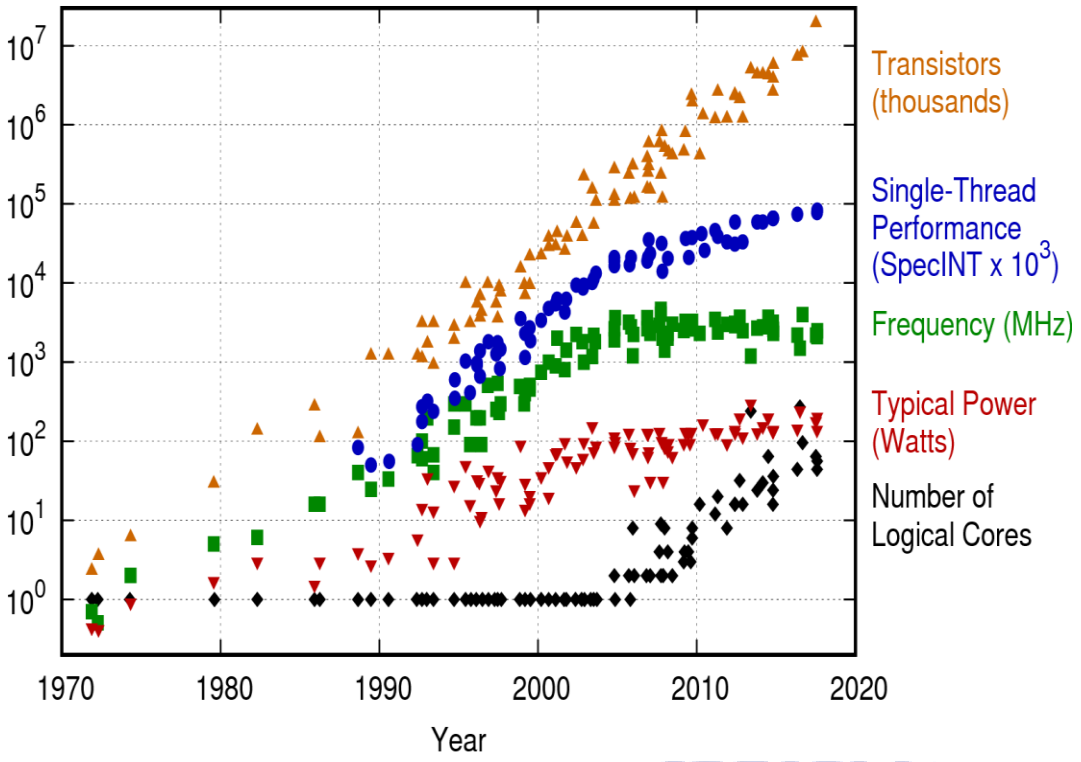
一是晶体管尺寸因不断逼近物理极限而减缓微缩。半个多世纪以来，晶体管尺寸微缩带动的性能提升、成本降低是推动集成电路制造工艺持续演进和计算技术指数级增长的重要驱动力。自工艺进入 10 纳米节点以后，晶体管性能随尺寸微缩提升幅度趋缓，主要表现在：晶体管集成度提升减缓，7 纳米节点单位晶体管面积缩小了 20%-30%，小于上一代的 37%；功耗优化减缓，7 纳米节点功耗降低约 10-25%，小于上一代的 30%；金属线宽的缩小导致阻抗上升，RC 延迟不利于芯片性能提高。此外，纳米节点制造工艺对光刻精度的要求快速提升，现阶段极紫外光刻（EUV）工艺仍不成熟。

二是因芯片过热而不可无限提升主频。依赖主频提升处理器单核性能是相对简单且高效的实现方法，但随着工艺尺寸稳步缩小，单位面积的晶体管数量翻倍增加，热累积效应愈加明显，处理器的功耗正以正比于主频的三次方量级高速增长。当处理器主频超过 4GHz 后，高频率下电子高迁移速率以及栅漏电现象导致产热量增加，使温度上升造成的性能损失超过主频对性能的提升，同时衍生重大的散热问题。受限于封装和降温成本的考虑，芯片主频的升级自 2005 年后即逐步

放缓，现大多控制在 4GHz 以内。

三是多核因并行算法局限而停止扩充。处理器自遭遇主频升级瓶颈后，开始转向多核架构，并通过增加并行计算能力以实现处理器性能的提升。经过十余年的发展，算法和软件的并行化依然不甚成熟，现有并行处理程序的编写、调试、优化能力仍然较弱，且大部分应用程序并不能自动分割任务交由多核处理，带来极大的软件重构和优化的工作量，致使处理器性能提升与核数不成正比，多核硬件的能力未得到充分发挥，实际应用水平远远低于理论能力。

四是冯诺依曼架构瓶颈日益凸显。冯诺依曼型计算机以“存储程序”为基础原理，程序执行时处理器在程序计数器的指引下顺序读取指令和数据，顺序执行形成计算结果。冯氏计算架构的特性决定了数据处理和数据读取二者速度需匹配方能保证计算的实时性和整体运行效率。随着摩尔定律的快速发展，处理器执行速度已远快于各级数据读取的速度，现阶段一级/二级缓存数据读取延迟 2-4 纳秒、内存延迟 70 纳秒、硬盘延迟 4 毫秒、外围存储介质延迟在秒级以上，数据读取与数据计算间的速度差异已成为制约计算效能升级的重要因素，冯诺依曼架构的瓶颈效应随着处理器计算速度的不断提升而更加凸显。多年来，高速内存、分支预测算法、哈佛架构等技术创新在一定程度上缓解了数据读取限流问题，但仍未实现本质改变。



数据来源：维基百科

图2 CPU 芯片计算能力发展历程（1970 年-2020 年）

2. 应用创新对计算需求的增速远超摩尔定律

数据总量激增，应用计算需求进入“新摩尔定律”时代。近年来，互联网、移动互联网、云计算、大数据、物联网、人工智能、5G 移动通信等 ICT 领域重大技术发展，加速推进社会迈入万物互联、万物感知、万物智能时代，逐步集聚和盘活海量数据资源。数据规模的增速远超摩尔定律，据 IDC 的数字宇宙报告，全球信息数据总量中接近 90%产生于近几年，据预测到 2020 年数据总量将达到 44ZB，平均个人拥有超过 5.2TB 数据规模。图灵奖获得者 JimGray 更是提出“新摩尔定律”，即每 18 个月全球新增信息量是计算机有史以来全部信息量的总和。数据结构趋于多元化，由传统文本等结构化数据扩展至图像、音频等不规则、非结构化数据，其中近三分之一的数据将具有大数据

开发价值，由此将带来极大的计算能力需求。

以人工智能为代表的算力依赖型应用极大加快计算资源消耗。除数据洪流催生计算资源和性能提升的普遍诉求外，以人工智能为典型代表的强算力消耗型应用创新更是极大提升了对计算能力的需求。区别于传统机器学习算法，以深度学习为代表的人工智能算法本质是基于概率统计理论，通过大量计算资源对大规模数据样本的处理，实现远超传统机器学习算法的识别精度，这种暴力计算模式正逐步流行并成为现阶段统治人工智能计算的主流范式。据统计自 2012 年以来，人工智能训练任务使用的计算能力每 3.5 个月提升一倍，目前增长已超过 30 万倍。大量计算资源的利用给算法、模型和应用的创新带来显著成果，如在 2012 年，谷歌与斯坦福大学组成的联合研究组利用 16000 台计算机处理数百万段 YouTube 视频，实现识别猫的功能。随着深度学习网络模型日益复杂、数据样本持续扩大，其对计算能力的需求和消耗与日俱增，人工智能计算体系已从早期的 CPU 芯片过渡到以并行处理性能取胜的 GPU 芯片，再到现阶段的大规模人工智能芯片集群，但与人工智能应用创新所带来计算需求增速相比差距依然较大。

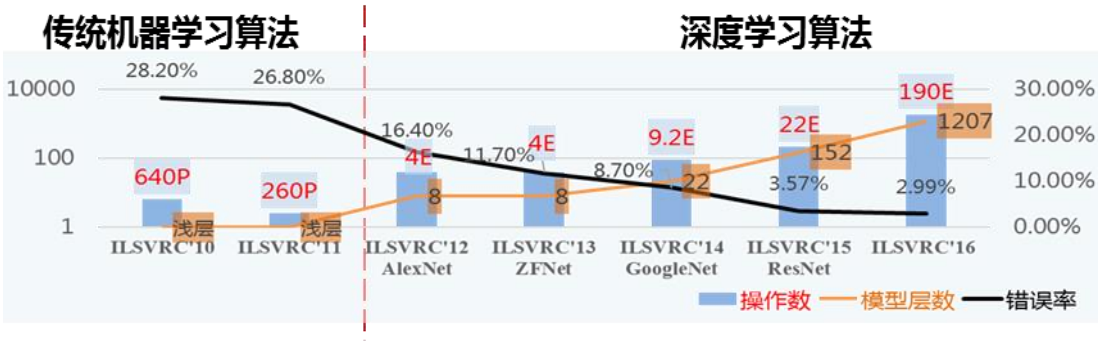


图3 传统机器学习算法与深度学习算法对计算需求的对比

（二）短期内需求驱动创新将主导计算技术产业的升级

1.发展动能转换：应用取代基础软硬件成为创新新动能

应用创新所带来的多样化需求成为计算发展的核心动能。计算的发展历程就是计算供给能力与应用创新需求之间的彼此驱动和迭代升级，二者的关系正由“先有能力，再谈需求”向“根据需求，实现能力”转变，创新应用在被动等待计算技术升级的基础上不断提升能动性，逐渐演变成为驱动计算发展的核心动能。大型机、小型机时代，由计算软件、计算硬件构成的计算设备/系统与应用基于一体化的模式发展，面向不同应用需求的软件和硬件均为专有体系，不仅昂贵且技术升级缓慢。PC 时代，以 IBM 代表的软硬一体化模式被英特尔和微软所打破，二者在软硬耦合的前提下，遵循“摩尔定律”和“安迪-比尔定律”滚动迭代，即计算芯片和存储器每 18-24 个月实现硬件性能的翻倍，以 windows 操作系统为代表的计算软件随之升级功能支撑应用创新。云计算及移动互联网时代，终端层面的智能手机和智能硬件在延续 PC 发展规律的同时进一步加快升级步伐，云端层面则通过虚拟化等软件技术实现大量计算硬件资源的汇聚以支撑搜索等应用的海量计算需求。目前，人工智能、自动驾驶、VR/AR 等创新应用爆发带来了计算需求的激增，现有计算硬件能力基本不能满足需求，差距普遍在十倍以上甚至百倍，传统计算升级模式已无法跟进应用快速创新的需求，计算进入应用直接定义的时代。

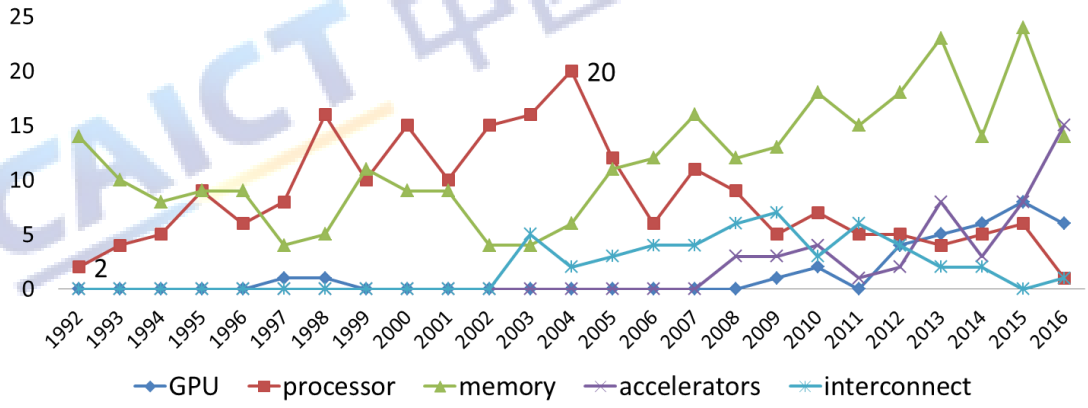
应用对计算的影响正在由系统软件向计算芯片不断深化。应用对计算的影响主要包括两方面：一是不直接影响计算硬件，通过计算软件实现对硬件资源的按需管理、灵活扩展，进而实现对上层应用创新

的支撑。此类模式的典型代表即为软件定义网络、软件定义数据中心、软件定义存储等一系列软件定义概念，其核心理念是通过虚拟化技术实现服务器、存储等计算资源的池化，通过对计算资源的统一化管理实现应用计算需求的按需分配。二是应用直接影响包括硬件和软件在内的整体化计算系统。目前人工智能、自动驾驶、VR/AR 等新兴应用对计算软硬件技术尤其是计算芯片创新的影响力日益凸显，以人工智能为例，本轮人工智能爆发是以深度学习为代表的算法突破为根本要素，算法本质是通过构建多隐层模型和处理海量数据提升识别准确率。深度学习算法区别于传统机器学习算法，以卷积、矩阵乘加等运算为主，海量数据搬运对芯片的并行计算能力、内存容量和 I/O 总线带宽等性能指标要求较高，擅长串行逻辑运算的 CPU 等传统通用计算芯片难以满足这种专用且高并行的计算需求，专用于深度学习计算加速的专用集成电路（ASIC）芯片快速崛起，据不完全统计目前我国从事人工智能 ASIC 芯片研发的企业就已超过 60 家。

2. 发展模式变换：以融合专用加速的系统优化为主

计算技术创新由通用软硬件到面向应用的专用加速。从近 20 多年来计算技术学术研究重点变化中可以看出计算技术创新重点方向已然发生变化。一是基础架构和基础工艺的技术创新日益趋缓。上世纪九十年代到本世纪初，是包括指令集架构 ISA、分支预测、超标量处理器及矢量处理单元等在内的处理器架构相关技术的活跃创新期，之后逐步趋于成熟、创新节奏放缓。以计算机体系结构国际研讨会 ISCA 发表论文为例，与处理器架构相关的论文 1992 年有 2 篇、2004

年有 20 篇，后逐年递减至 2016 年的 1 篇。二是存储架构和互联架构因多核及分布式的需求愈发重要，自本世纪起步以来贯穿至今。在处理器架构技术创新趋缓的同时，因冯诺依曼体系所带来的 I/O 瓶颈优化及多核、分布式体系的升级需求，存储架构和互联架构的创新自本世纪初日益得到关注，同期 ISCA 所发布的技术成果均达到近 20 年高峰，后期更是成为业界科研关注的重点方向，并成立专门的国际研讨会进行技术创新成果的发布。三是部分通用及特定专用的加速架构逐渐兴起，包括高性能计算、智能终端计算以及人工智能计算等新兴应用领域计算需求的激增，尤其深度学习的爆发是近年快速增长的核心驱动力，推动 GPU 加速计算及专用 ASIC 加速计算等相关技术创新异军突起，成为计算技术创新的主导方向。



数据来源：《技术驱动架构创新：过去，现在和未来》（UCSB，谢源）

图4 计算机体系结构国际研讨会 ISCA 论文分类数量统计（1992-2016）

处理、存储和互联等三大部件技术和计算系统技术均围绕专用加速开启密集创新。目前面向实际应用场景需求的计算技术升级均为冯诺依曼体系内的技术创新，主要体现在三方面：一是数据处理方面，正逐渐由以实现逻辑控制和通用计算的 CPU 处理器构成计算平台，演

变为 CPU 与 GPU、FPGA、DSP、各类深度学习加速 ASIC 等具备专用计算能力的硬件相结合，构成可覆盖多源数据多样处理需求的混合加速硬件平台。二是数据存储方面，现有缓存、内存及硬盘等各级存储介质通过设计技术和工艺技术的升级不断提升存储密度和存取速度，与此同时高速非易失性内存 NVM（Non Volatile Memory)等新兴存储介质技术也在不断发展，并凭借接近系统内存的读写性能以及与硬盘类似的非易失性特点实现对现有多级存储架构的重构。三是数据交互方面，主要围绕高速和共享两大方向升级，在包括 PCIe5、Nvlink、NVSwith 等总线技术以及 25G 以太网技术等板级和系统级互联技术不断高速化升级之外，多种互联技术均强调优化计算单元间的共享数据访问，尤其是 CPU、GPU、ASIC 等多样处理单元间的内存一致性访问，以加快计算单元与存储单元件间的数据交互、缓解冯诺依曼 I/O 瓶颈限制。在上述三大计算要素并行创新的基础上，面向不同应用场景的差异化计算需求，通过计算系统技术的协同创新以实现整体系统在计算性能、功耗、延迟等方面的平衡高效也成为后续升级的重中之重。

三、 现阶段先进计算创新重点

（一） 器件技术：多路径推动摩尔定律持续演进

摩尔定律仍在延续，2018 年制造工艺全面升级。台积电、三星、英特尔三大巨头持续推动先进工艺研发及规模应用。一是目前已全面进入 10/7 纳米工艺节点。台积电 2018 年 4 月量产 7 纳米工艺，相比上一代 10 纳米工艺，芯片功耗降低 40%，性能提升 15%，核心面积缩小 37%，目前华为麒麟 980、苹果 A12 均采用该工艺实现量产。英特

尔最新 10 纳米工艺的晶体管密度可达到 100.8M/mm^2，仍然是先进工艺的最有力竞争者，计划 2019 年投入量产。二是 EUV 等新技术迈入应用阶段。台积电于 2018 年 10 月完成了首次 7 纳米 EUV 流片，相比初代 7 纳米工艺性能提升 15%以上。三星也已完成 7 纳米 EUV 工艺的研发，并同时宣布了技术发展路线图，将 3 纳米节点提上研发日程。

尺寸微缩逼近物理极限，升级难度日益加大。2018 年先进工艺阵营再次减员，全球第四大代工厂格罗方德宣布放弃 7 纳米及以下节点技术研发，将资源集中在现有的 14/12 纳米制程产品；同期台湾联电也宣布停止 12 纳米以下工艺的研发。目前，全球仍有 7 纳米及以下节点研发计划的仅剩台积电、三星、英特尔三家企业。作为集成电路行业标杆的英特尔也已废止了两年一循环的“制程-架构”产品研发周期，自 10 纳米后开始采取“制程-架构-优化”三年三步走策略。

晶体管结构创新加速，推进芯片制造工艺能力升级。晶体管技术创新从未停止，从 90 纳米到 10 纳米先后经历了引入应力、加入高 κ 栅介质、采用鳍式场效应晶体管 FinFET 结构以及改变栅极接触位置等创新材料/技术的应用。目前已有多家厂商开始针对 5 纳米及以下节点工艺制程的晶体管结构进行研发，IBM 和三星分别针对 5 纳米和 3 纳米工艺提出了 Nanosheet 和 MBCFET 结构，二者的优势在于通过构建多沟道环栅结构，使晶体管的电流驱动能力以及栅极对载流子的控制能力得以提升，预计将于 2024 年代替 FinFET 结构成为主流。此外，更为多样化的晶体管结构创新仍在不断探索中，采用Ⅲ-Ⅴ族化合物半导体作为晶体管沟道材料，可提升电子迁移率，使晶体管获得

更大的电流驱动能力；新型隧穿晶体管（TFET）利用载流子隧穿原理实现超陡亚阈摆幅，达到降低晶体管功耗，提升能效比的效果；垂直纳米线结构可进一步减小单个晶体管所占面积，大幅提升集成度，推动制造工艺迈向更小节点。

三维堆叠提升集成密度，等效延续摩尔定律。目前三维结构在存储领域已经有所应用，三维存储结构 3D NAND 通过增加存储叠层而非缩小单个存储单元的尺寸实现了存储密度的增长，解决了传统二维半导体存储芯片中存储单元不断缩小导致的成本上升以及相邻存储单元之间的串扰问题，成为未来实现存储芯片容量可持续增长的关键技术。2018 年，三星、东芝、美光、海力士等存储器厂商先后发布 96 层 3D NAND 存储器产品，存储密度达到 4Gb/mm^{2}以上，与 16nm 工艺条件下二维 NAND 存储器相比，存储密度提高 4 倍以上。据最新国际半导体技术路线图（ITRS2.0）预测，未来三维叠层结构还将在多功能复合芯片等领域发挥关键作用，复合芯片将聚合传感器、新兴存储器和硅基电路，在一颗芯片上实现信息采集、存储、计算和输出等功能。

系统级设计和多质多维封装同步深化，加速芯片多功能集成创新。通过面向更多功能需求的设计及封装技术，以进一步提高芯片集成度、降低整体功耗、推动多功能异构的发展。芯片设计方面，采用片上系统（SoC）设计方式实现各关键功能部件的片上集成，达到降低功耗、减小电路面积、提高系统各部件之间通信速度的效果，目前片上系统设计技术已非常成熟，是移动互联网、物联网等领域芯片设计的主流思路。先进封装技术方面，通过将多个功能芯片通过封装技术以达到

提高芯片集成度的效果，可分为系统封装（SiP）、3D 堆叠封装以及一体化 3D 封装，其中：SiP 封装可将各种工艺下、不同种类的芯片进行集成，该技术开发周期短、成本低；3D 堆叠封装将多个芯片按垂直堆叠的方式进行封装，并利用垂直通孔（TSV）技术实现层间连接，相比系统封装可进一步提高芯片的集成度，但目前较为成熟的工艺仅能通过金丝球焊以及焊接球的方式实现两层电路之间的连接，对于三层以上电路的封装，工艺尚未成熟；一体化 3D 封装是采用更为密集的垂直互连方式将位于各层的晶体管按照设计规则相连组成功能电路，以最大限度利用垂直维度，达到节省芯片面积的目的，该方法对集成电路制造工艺要求较高，目前仍处于实验室阶段，进入商用尚需时日。

（二） 部件技术：三大计算单元加速协同创新

从上世纪五十年代的第一台冯氏结构计算机 ENIAC 到今天，计算设备和计算系统在外观形态、部署方式、应用特性等方面虽发生了翻天覆地的变化，但体系结构依然遵从冯诺依曼架构，计算设备的主要组成部件以及彼此之间的交互机制也相对稳定。除键盘、鼠标等输入设备以及显示器、打印机等输入设备外，与数据处理相关的运算器和控制器、与数据存储相关的各类存储模块、以及数据在上述两大单元间实现交互的通信类接口和模块是构成计算设备和计算系统的主要功能模块，也是构成计算技术体系的三大重点单元。

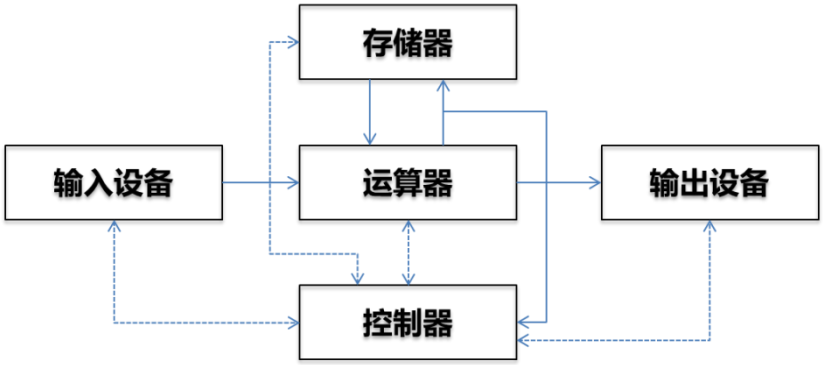


图5 冯诺依曼架构图

1.数据处理单元

CPU、GPU、FPGA 是目前通用计算领域的三大主流计算芯片。CPU 芯片兼顾控制和计算，是构成笔记本、智能终端及服务器计算硬件主体。CPU 芯片架构中“缓存-控制-计算”三者兼顾，其中 70%晶体管作为缓存和控制单元，控制单元用于实现分支预测、流水线等复杂逻辑等，大量缓存单元降低数据读取时间以降低延时；30%晶体管作为计算单元，可在 1-3 个时钟周期内完成双精度浮点乘加等运算操作。**GPU 芯片适合通用并行处理，应用领域由早期图像处理逐步拓展至通用加速。**GPU 芯片内置大量计算单元和高速缓存，逻辑控制单元相对简单。控制单元负责将串行计算任务拆分成多个相对简单的并行任务，缓存单元配合进行数据高速转发，最终由大量的计算单元完成分拆任务的计算并返回最终结果。因而 GPU 架构适合逻辑相对简单的大批量高密度同构计算任务，但因密集执行计算会带来大量功耗和发热。**FPGA 芯片具备可重构特性，可根据客户需求灵活定制计算架构，更适合于需求量偏少的航空航天、车载、工业等细分行业。**FPGA 芯片具备可编程的数字基本门电路，可依据应用灵活配置底层架构，由于无需指令和软件控制，直接采取晶体管电路实现应用算法，相比于

CPU、GPU 芯片响应更快，更适合于流式计算密集型任务。但 FPGA 芯片编程要使用专门的硬件描述语言，技术门槛相对较高，大规模应用也不具备成本优势。

三大计算芯片技术创新依然活跃。一方面持续挖掘传统架构技术潜力。CPU 不断优化现有架构技术能力，采用乱序执行、超标量流水线、多级缓存等技术提升整体性能表现；GPU 持续探索高效的图形处理单元、流处理单元和访存存取体系等，并优化编程框架降低 GPU 编程和应用程序移植难度；FPGA 不断强化应用功能的丰富完善，升级芯片内部组件以适应广泛的加速场景，并发展基于 C/C++、OpenCL 等软件工具开发生态，降低开发者门槛。**另一方面均通过引入专用计算能力迎合人工智能等新兴领域的计算需求。**当前，受摩尔定律发展趋缓以及功耗墙限制等影响，单纯依赖升级制造工艺、增加处理器核数、提高时钟频率等传统升级路线带来的性能提升幅度有限，业界开始转变技术发展思路，借助专用计算单元提升专用领域的适用性。CPU 围绕深度学习计算需求增加专用计算指令，如 ARM 公司最新 Cortex-A76 芯片通过优化缩减深度学习常见的乘法累加运算周期等，实现边缘侧人工智能性能相较于上一代产品提升接近 4 倍；GPU 持续优化针对人工智能计算的专用逻辑运算单元，英伟达图灵架构 GPU 芯片内置全新张量计算核心，利用深度学习算法消除低分辨率渲染问题；FPGA 提升面向各类工作负载需求的异构计算能力，以实现边缘智能等更多场景的规模应用。

AI ASIC 现已成为专用计算加速芯片创新的典型代表。专用集成

电路（ASIC）意指针对特定领域、特定算法需求设计的电路，与通用芯片相比面积小、性能高、功耗低，大规模量产后具备成本优势，可广泛应用在市场需求量大的专用领域。**目前为满足人工智能应用计算需求的 AI ASIC 是创新的焦点所在，升级重点围绕指令集、计算架构、访存体系、交互通信等四大方面。**因以 CPU 为代表的传统通用芯片在基础能力上无法满足深度学习等人工智能复杂神经网络算法、海量数据高吞吐量、高密度线性代数任务的处理需求，ASIC 芯片通过软硬融合的极致性能以及大规模量产低成本等优势，正成为业内创新重点，围绕深度学习指令集、高并行计算架构、高能效访存架构、高速低延时互联等持续升级。指令集方面，主要针对深度学习算法中高频、高耗时的矩阵、向量等逻辑运算进行优化，并简化与算法无关的分支跳转、缓存控制等逻辑控制指令。计算架构方面，多选择众核等高并行架构进行设计，并集成矩阵乘加等专用运算单元增强针对深度学习算法共性计算需求的支撑能力。如谷歌将张量处理单元（TPU）引入脉动阵列架构，实现数据高效复用功能，提升并行处理能力；集成超过 6 万个计算核心单元组成专用矩阵乘加模块，提升深度学习算法计算效率。存储方面，应用高带宽内存等新型技术提升内存带宽，结合片上内存、数据复用、模型压缩等手段降低内存存取频次。如寒武纪公司早期学术论文中即提出可采用大量的片上存储设计来降低片外存储访问需求和数据访问功耗，这一设计方案目前也被谷歌等众多企业广泛采用。互联方面，配置新型 PCIe 5、CCIX 等高速易扩展的异构互联总线，通过带宽加速缓解海量数据频繁读取所导致的高延时问题。

结合场景需求和算法特征定向优化，AI ASIC 芯片差异化创新加速。由于人工智能的不同应用场景间差异性较大，难以通过一款通用人工智能芯片适合所有领域，随着各应用场景定位和需求的逐步明确，AI ASIC 呈现多技术路线分化态势。深度学习计算主要分为训练和推理两个阶段。其中，深度学习模型训练以高性能、高精度、通用化的计算能力为主，芯片需堆叠大量高精度浮点运算单元、高带宽内存和专用计算单元等提升训练效率，但受限于高能耗目前多集中在云端部署实施。推理阶段则因应用场景的不同而各具差异，云端推理芯片多应用低位宽定点运算单元、片上内存等实现高通量、低延时、通用化的推理能力；端侧推理芯片则需要深度耦合特定场景和神经网络算法，利用低位宽低精度运算、模型压缩等技术实现低时延或低功耗等差异化场景需求。以谷歌 TPU 系列芯片为例，面向云端训练的 Cloud TPU 芯片采用高精度的 32 位浮点和标量运算器，封装 16GB 高带宽内存，可实现每秒 180 万次的峰值浮点操作能力；面向云端推理的 TPU 芯片采用相对高精度的 16 位定点运算器，集成 24MB 的片上内存，可实现每秒 92 万亿次的峰值定点运算性能；面向端侧物联网领域的 Edge TPU 芯片采用 16 位和 8 位定点运算单元，每秒可完成 30 帧高分辨率视频处理，具备面积小、功耗低等优势，可广泛应用于边缘侧推理任务。

表1 不同应用领域对 AI 芯片的需求对比

	云侧训练	云侧推理	端侧推理
性能需求	高性能 高精度 通用性	高通量 低延时 通用性	高能效 低延时/低功耗
关键技术	高并行计算架构	高并行计算架构	高并行计算架构
	专用矩阵乘加单元，兼具标量处理能力	专用矩阵乘加单元，兼具标量处理能力	张量计算单元
	高位宽、高精度、浮点运算	高位宽、高精度、浮点运算	低位宽、低精度定点运算
	高带宽内存 HBM	大片上内存 SRAM	通过模型压缩、数据复用，减少内存访问
	高带宽易扩展接口，如 Nvlink/PCIe5 等	高带宽易扩展接口，如 Nvlink/PCIe5 等	高速、丰富的处理器接口，如 MIPI 等

2.数据存储单元

数据存储技术发展迅速，现有体系不断演进。自“存储程序”的计算机体系确立起来，存储单元始终与数据处理单元同步升级。当前，互联网、移动互联网、人工智能浪潮迭起，数据信息爆发式增长，对存储器性能和容量的需求与日俱增，存储器技术体系也在不断变革。半导体存储器根据是否需要供电实现数据保存可分为易失性存储器和非易失性存储器，其中，易失性存储器以 DRAM 内存为代表，广泛应用于移动设备和服务器设备中，并通过架构创新、TSV 先进封装等不断提升存储密度、降低延时；非易失性存储器以 NAND 闪存为代表，以其为介质的固态硬盘正替代传统机械硬盘，成为数据中心的主流存储器件，现阶段围绕 3D 集成、层数堆叠等方向升级存储容量。此外，相变存储器、磁阻式存储器、阻变式存储器、磁畴壁存储器等新型非易失性存储器，因兼具密度高、功耗低、读写快、反复操作耐受力强等优势，备受业界关注。

表2 三种典型的新型非易失性内存

	PCM	阻变存储器	磁阻存储器
实现原理	利用特殊材料在晶态和非晶态之间相互转化时所表现出来的导电性差异来存储数据	由外部刺激引起存储介质离子运动和局部结构变化，造成电阻变化，并利用这种电阻差异来存储数据	靠磁场极化而非电荷来存储数据，通过判断两层磁介质极化方向来读取存储数据。
技术现状	英特尔与美光合作生产 3DXpoint，已开始小规模应用	Crossbar、闪迪、东芝等公司已研制出 RRAM 测试芯片	Everspin 已推出 MRAM 产品，三星等在积极研制
应用瓶颈	需进一步提高存储密度、降低成本和提高耐写能力	需解决存储单元一致性以及存储单元之间的串扰问题	需要解决相邻单元之间较为严重的串扰问题
参与企业	英特尔、美光	Crossbar、东芝、闪迪、索尼、三星	Everspin、三星、IBM

非易失性内存等新型存储技术创新活跃，或将变革现有存储层次结构。现有计算设备和系统中数据存储单元是由片上缓存、片外内存、固态硬盘等构成的多级存储架构，通过多级存储弥补处理器和存储器间巨大的速度差异，但一旦需跨多级读取数据就会带来大量的数据传输层数和 I/O 调度的额外开销。当前，以英特尔的 3D Xpoint 为代表的新型非易失性内存引发了业内极大的关注，其大规模应用后将对现有多级存储体系架构带来颠覆性影响。新型非易失性内存的优势主要体现在两方面：一是具备按字节访问、支持虚拟地址读写等特性，可减少 I/O 管理调度开销，二是容量大、访问速度快，可直接连接在内存总线，替代原有内存和固态硬盘的位置，减少现有存储层次结构，文件读写性能相较于现有存储结构提升数十至数千倍，同步也将带来文件系统、I/O 软件栈等系统软件技术和数据库技术的变革。

3.数据交换单元

高带宽和直接内存访问是目前缓解冯诺依曼瓶颈、提升计算效率的重要创新方向。除了提升总线和以太网交换速度等传统升级手段外，板卡级和系统级数据交换分别以内存共享、远程直接内存访问（RDMA）等为方向提升内存存取性能。**板卡级层面，**通过引入统一的虚拟内存（UVA）和 NVLink 高速总线技术，可实现“CPU+GPU”异构并行计算中跨节点内存操作，以缓解 GPU 因单节点本地内存不足而降低并行效率的问题：一方面 UVA 允许多个 GPU 节点之间合并共享彼此的显存空间，并允许 GPU 直接访问并利用系统内存，另一方面从 PCIe 总线升级到 NVLink 总线，提升内存空间数据传输速度，可实现数据交换总体效率提升接近 10 倍。**系统级层面，**传统的 TCP/IP 网络通信是通过内核发送消息，并经过一系列多层网络协议的数据包处理工作，由此会带来高额的 I/O 处理开销，限制了跨服务器间传输的带宽及灵活性。RDMA 技术在无需操作系统介入情况下，实现数据从系统内存直接传输到远程系统存储器，消除了外部存储器复制和上下文切换的开销，实现了高吞吐、低延时的网络数据通信，目前 InfiniBand、iWARP、RoCE 等常见高速互联传输技术均支持 RDMA 技术。

（三） 系统技术：围绕应用需求展开体系化升级

计算技术发展到今天，其承载的物理实体早已超越早期大型机、小型机、PC 计算机等单台设备的概念，逐步走向计算系统的发展理念。目前牛津大学计算词典中对“Computer”的解释为能够以明确定义的方式执行一系列操作的设备或系统。现阶段包括云计算、边缘计

算、异构计算、高性能计算等层出不穷的计算概念也均是从系统角度出发。因此，在本研究报告中，不对桌面电脑、笔记本电脑、服务器、智能手机等简单计算设备做过多的阐述，而是从不同计算系统的特性出发，围绕目前业界较为活跃的三类计算系统中所涉及到的关键技术进行分析研究。主要包括三大方向：从计算资源种类角度出发衍生而来的异构计算和可重构计算、从计算聚集模式角度出发衍生而来的分布式计算和集群计算、从三大计算单元协同发展衍生而来的内存计算和存算一体化。

1.异构及可重构

异构计算硬件先行，软件作用日趋凸显。随着大数据时代的到来，异构计算已成为突破计算能力和功耗瓶颈的有效途径之一。异构计算共历经四大发展阶段，分别是单纯挖掘并行潜力、添加专用加速单元、针对特定应用领域定制、多种平台的高效融合。**目前异构计算的发展重心已经从硬件开发转移到深化应用、软硬件融合创新阶段，软件对异构计算的支撑作用越来越明显。**计算系统会应用到多种硬件体系结构，例如搜索、解析等有大量控制代码的程序主要运行在 CPU 上；大量数据处理的程序适合运行在 GPU、DSP 等矢量体系结构处理芯片上；大量专用计算的应用适合在 FPGA、ASIC 等针对应用进行优化的芯片上执行。为了提高程序执行性能，需要 OpenCL 等标准化的异构编程框架，来完成跨平台的并行编程任务，从而实现对底层多种硬件平台的高效利用。

CPU+GPU、CPU+MIC、CPU+FPGA 是目前三大主流的异构计算技术。

CPU+GPU 计算平台是目前比较成熟的异构平台，可满足高并行和高吞吐的计算需求，适用于计算密集、高并行、SIMD 应用，擅长图形图像、矩阵计算以及深度学习相关应用。MIC（Many Integrated Core）架构由英特尔推出，是将数十个精简的 x86 核心整合在一起的处理器，相比较 CPU 而言具有更多的硬件线程和更宽的矢量单元，可更好满足高并行度应用需求，其在计算机体系中作为协处理器存在，并具有与 CPU 相同的编程模式、语言和优化方法，降低了软件编程的难度。FPGA 异构计算则具有更高的效能，并且并行模式更宽泛，支持数据并行和任务并行，计算密度和灵活性较高。

端侧异构已较为普及，云侧成为下阶段发展的重点。异构计算在终端领域已得到广泛应用，如高通骁龙移动平台集成了 Hexagon DSP 向量处理器、Adreno GPU 视觉处理器和 Kryo CPU 的多核异构计算核心，面向不同类型的功能、基于不同类型的数据、在不同的计算精度水平上，提供更高效的计算解决方案；华为推出的麒麟 980 芯片同样采用基于 CPU、GPU、NPU 等全系统融合优化的异构架构，并集成了双核 NPU 神经网络处理单元，图像识别速度比 970 提升 120%。云侧异构不断深化，不断变革传统云计算硬件基础架构体系。GPU 异构正成为实现云端计算平台的重要技术模式，目前全球 TOP500 HPC 中有 100 家使用 GPU 加速，96%采用了英伟达的 Tesla 系列 GPU 产品；百度、谷歌、Facebook 等超大规模云计算中也部署了 GPU 密集计算能力。利用 FPGA、ASIC 等提升云端异构计算能力仍在不断深化，如英特尔已提供整合 Altera 的异构平台方案，赛灵思也与 IBM 合作加

速布局，预计 2020 年 1/3 的云服务商将采用 FPGA；谷歌 TPU 开放 1000 个云 TPU 服务集群，每个云 TPU 最高可达 180TFLOP/s 浮点计算，配备 64GB 带宽内存。

数据中心异构体系开启由 GPU 到 FPGA 的新变革。FPGA 并行计算能效比（矩阵、信号处理等）是 CPU 的 33 倍、GPU 的 3 倍，并同样适合串行加速典型场景（如 I/O 操作等），因此对数据中心异构计算的变革性作用愈为凸显。目前主要存在三种场景的应用探索：一是专用 FPGA 集群模式，主要是指采用专用 FPGA 设备组成的集群来实现对数据中心计算任务的分摊和卸载，其应用的局限性在于不同服务器的 FPGA 间无法通信，加速规模受限于单台服务器集成 FPGA 数量，网络延迟无法保证，专用机柜单点故障影响加速，并且需专用定制服务器。二是单机 FPGA 专用网络模式，即数据中心的每台机器都配备一块 FPGA，并采用专用网络连接。三是单机 FPGA 共享网络模式，与第二种的主要区别在于异构部署融入数据中心网络中，FPGA 部署在 NIC 网卡和以太网交换机之间，虚拟交换机的数据平面功能被转移到 FPGA，释放对 CPU 的计算资源占用。现阶段 FPGA 专用加速应用较为广泛，代表厂商如微软、亚马逊、Facebook、百度、阿里、腾讯等互联网企业等。

全球企业加速异构计算的布局。随着摩尔定律接近物理极限，仅提升单一处理性能的代价越来越高，横向拓展异构计算变得愈发重要，目前已经成为众多厂商的主流发展方向。英伟达于 2018 年 6 月发布机器人异构计算平台 Jetson Xavier，集成 6 种高性能处理器，

包括 1 个 Volta Tensor Core GPU、1 个 8 核 ARM64 CPU、2 个 NVDLA 深度学习加速器、1 个图像处理器、1 个视觉处理器和 1 个视频处理器。英特尔 AI 计算平台产品体系也囊括了 CPU、GPU、DSP、NNP、FPGA 等不同处理核心，覆盖云端数据中心到设备终端。此外高通、三星、华为、苹果等移动芯片企业也在不断加速 CPU+GPU+DSP+xPU 的移动异构计算平台升级。

可重构计算可灵活适配不断变化的计算需求。经典计算架构一旦确定即不可改变，应用需根据计算结构的特性进行优化以实现最优的效能表现。可重构计算是一种函数化的硬件架构，允许系统硬件架构和功能随软件变化而变化，以满足软件不断变化的计算需求。可重构计算并非全新概念，上世纪六十年代加州大学洛杉矶分校的 Gerald Estrin 教授即提出：计算机可以通过一个主处理器加上一组可重构硬件组成，主处理器负责控制可重构硬件的行为，可重构硬件根据任务的计算特点，通过剪裁、重组等方式，达到加速执行某一特定任务的目标。随着集成电路工艺的不断演进，上世纪九十年代后，可重构计算相关的软硬件技术研发日益高涨，目前已有 PACT 公司的 XPP-III、IPFlex 公司的 DAPDNA-2、IMEC 的 ADRES、Freescale 公司的 MRC6011 等多款可重构计算产品。此外，斯坦福大学、加州大学伯克利分校、卡内基梅隆大学、以及国内的清华大学等高校也正在对可重构计算进行深入研究。

人工智能成为可重构计算技术发展的重要驱动力。当前 AI 正在加速计算芯片的架构创新，对于现在尚未定型的各类 AI 算法而言，

可重构计算成了 AI 芯片设计的一个重要研究方向。清华微电子所可重构计算团队推出了代号为 Thinker 的系列芯片，其中 Thinker 1 为一款实验性质的验证芯片，证明了“软件定义芯片”在 AI 芯片设计中的可行性，此项技术在 2017 年 ACM/IEEE ISLPED 国际低功耗电子学与设计会议上获得设计竞赛奖；Thinker 2 为一款人脸识别芯片，可做到 6ms 人脸识别、准确率超过 98%；Thinker S 为一款语音识别芯片，功耗只有 200 多微瓦，且可以进行声纹识别。2017 年 12 月澜起科技携手联想推出了基于该可重构技术的津逮服务器 CPU 及平台，标志着我国可重构计算技术取得重大突破。此外赛灵思、英特尔等国外厂商也在加速可重构计算技术的研发布局。

2.分布式及集群

分布式计算是一种共享软硬件资源的计算形式。分布式计算是指将一个大型计算任务分解，由通过网络互联的若干计算系统分别处理，最后将所有的计算结果合并为原问题的解决方案。分布式计算主要经历了三个阶段：第一阶段是 1980 年到 1990 年，主要是通过通信互联的创新实现多设备间的信息共享。第二阶段是 1990 年到 2000 年，面向对象技术的迅猛发展，急需解决大量应用之间的互操作问题。基于分布式对象、中间件等软件技术创新活跃，大型分布式计算系统得以快速发展。第三阶段是 2000 年后至今，互联网应用的融合创新使得多业务层面跨系统协作的需求日益增多，基于分布式计算实现的 Web 服务和 SOA 技术（服务计算）等成为发展重点方向。

Hadoop、Spark 和 Storm 是目前最重要的三大分布式计算系统。

三者作为目前较为主流的分布式计算框架，有各自的适用场景，其中 Hadoop 通过将数据切片分别计算来处理大量的离线数据，常用于海量数据的离线分析处理；Spark 是一个基于内存计算的开源集群计算系统，运算速度超过 Hadoop100 倍，但不能用于处理需长期保存数据，常用于对离线数据的快速分析；Storm 侧重流式计算，不进行数据收集和存储，通过网络实时接收和处理数据，实时传回结果，可实现对大数据流的实时处理，常用于在线实时的大数据处理。

分布式计算正由中心能力集聚向边缘融合扩充发展。多数云数据中心是集中化的，离终端设备和用户比较远，对于实时性要求高的计算服务通常会引起长距离往返延时、网络拥塞、服务质量下降等问题。而边缘计算是将数据在边缘网络进行本地化处理，包括终端设备、边缘设备、边缘服务器等，强调计算的去中心化/去本地化部署，计算服务需求响应更快。目前在计算技术中出现了多种云和边缘计算模式，移动云计算（MCC）和移动边缘计算（MEC）已成为云计算和边缘计算的扩充。其中 MCC 在边缘网络中提供轻量级的云服务器，支撑靠近终端用户的移动应用程序在远程执行。MEC 是移动通信演变的关键因素，目前已成为 5G 网络架构的重要组成部分，其将边缘服务器和基站相结合，与远程云数据中心互联，为用户带来自适应和更快初始化的蜂窝网络服务，提高网络效率。

集群计算侧重合并多计算平台实现同一任务。集群计算是指由众多计算机软件或硬件构成一个紧密协作的计算主体，以实现对计算任务的高效处理。从组成集群系统的计算机体系结构的角度，可将集

群系统分为同构与异构两种；按功能和结构的差异性，可分为高可用性集群（Highavailability Clusters）、负载均衡集群（Loadbalancing Clusters）、高性能计算集群（Highperformance Clusters）、网格计算（Grid Computing），其中高可用集群主要用于保护所承载应用的连续不间断，降低因各种故障对业务的影响；负载均衡集群一般通过一个或者多个前端负载均衡器，将计算任务分发到后端的一组服务器上，以保证整个系统的高性能和高可用性；高性能计算集群采用将计算任务分配到集群内不同计算节点以提高计算能力，主要应用在科学计算领域。

高性能计算（超算）已成为集群计算的重要应用领域。HPC 是利用并行处理和互联技术将多个计算节点连接起来，从而高效、可靠、快速地运行高级应用程序的过程，多被称为超级计算，已成为解决科学研究、经济发展、国家安全等方面诸多重大难题的重要手段。超级计算机性能的发展遵循千倍定律，即每隔 10 年性能就会提高近千倍，如 2008 年 IBM 推出历时 6 年研制的走鹃超级计算机，最大运算速度为 1.015PFlops；2018 年 IBM 推出 Summit 超级计算机，每秒高达 20 亿亿次(200PFlops)的浮点运算速度峰值。经过近九十年的发展，超级计算已从追求性能为主、应用局限在科学与工程计算、只有少数企业参与的初级阶段，发展到通过低成本短时间提升性能、应用扩大到互联网和企业数据中心等领域、形成了较为完备的高性能计算产业的中级阶段，未来高性能计算将进入效率优先的时代，形成核心部件、互联网络、编程框架等协同创新的体系。

P 级超算时代 GPU 等异构式超算开始崭露头角。随着软件的快速配套和并行集群计算技术的加速发展，CPU+GPU、CPU+MIC 众核等异构式超算已逐渐走向成熟，有助于提升性能功耗比，为包括大量矩阵计算、卷积计算的算法提供算法加速。目前全球几乎所有高性能超算系统都采用异构架构，其中 Summit、天河 1 号、泰坦、代恩特峰等属于 CPU+GPU 异构式超算；天河 2 号、科里、Oakforest-PACS 采用 CPU+MIC 众核加速器，也属于异构超算体系；神威太湖之光虽然只采用了一种申威 26010 众核处理器，但因申威处理器内集成了 4 个管理核心和 256 个运算核心，也可隶属于异构超算。

中美日欧加速推进，2021-2023 年将先后进入 E 级计算时代。E 级计算特指功耗在 20MW-40MW 内，计算速度超过每秒百亿亿次浮点运算，将是超级计算发展的又一里程碑，中国、美国、欧洲和日本是主要推动的国家。美国能源部主导的 E 级计算计划（ECP）累计投资 4.3 亿美元，计划 2021 年至 2023 年推动 3 台 E 级系统分步上线；日本作为最早明确其 E 级计算发展路线的国家，计划 2021 年推出基于 ARM 架构的 E 级计算机 Post-K Computer；欧盟主导的 EuroHPC 计划已有 20 个欧盟国家加入，欧盟政府将投入 4.8 亿欧元，通过与各国政府共同投入的模式，在 2020 年建设两台 Pre-E 级系统和两台 P 级系统，并在 2022-2023 年左右建成 2 台 E 级系统，且其中至少有一台将使用欧洲自主的技术；我国的 E 级超算系统研制将继续依托科技部重点研发计划高性能计算专项开展，由国防科大、江南计算机研究所和中科曙光等 3 家单位分别开展 E 级系统技术研发和验证，计划 2020 年完

成研制。据目前的技术路径而言，E 级计算虽可基于 P 级计算升级扩展实现，但若希望得到更优效能表现，面临访存、通信、可靠性、能耗和可扩展性等五大方面的挑战，且围绕化学与材料科学、能源开发与利用、地球与空间科学、数据分析与优化等多个方向的应用成为实现 E 级计算生态系统构建的关键。

3. 内存计算及存算一体化

内存计算成为突破冯氏体系瓶颈制约、提升整体计算效能的有效举措。传统的冯诺依曼架构体系受限于“存储墙”，即计算速度和数据读取速度之间巨大的鸿沟，基于内存计算理念实现数据就近计算成为提升性能的最直接有效方式，目前围绕此方向存在三种发展思路：内存内计算（in-memory computing）、内存驱动计算（memory-driven computing）和存算一体化（processing-in-memory）。内存内计算主要通过数据库等软件技术实现内存数据直接读取，并进行实时处理和分析，是对传统数据处理方式的一种加速。内存驱动型计算是打造多个处理器共享同一内存池的系统体系，通过高速、低功耗的内存互联架构实现灵活扩展和效率提升，如 HP 原型机 the machine。存算一体化目前有两种发展路线：一是在内存和固态硬盘芯片中植入逻辑计算单元，适合云端大数据类应用和神经网络训练等；二是使用 NOR 等存储器件单位直接完成计算，适合终端设备的 AI 推理等。

内存内计算是目前最主流的内存计算方式。在摩尔定律的推动下，内存容量提升、读取速度加快和价格持续下降是推动内存内计算技术日益普及的主要因素。伴随数据量的增加和计算复杂度的提升，

内存内计算技术已历经分布式缓存、内存数据网格、分布式内存数据库和高性能、集成化、分布式内存平台等四大发展阶段。内存内计算在硬件上主要采用多核处理器架构和以 TB 计的内存服务器，软件创新则主要聚焦在内存数据库，采用列式存储机制大量降低系统 I/O，并集成高效的数据压缩、动态聚合等技术，优化内存利用，以实现数据的密集、实时运算。

SAP HANA 和 Apache Spark 分别是当前较主流的商业和开源内存内计算技术。SAP HANA 是一项在本地内存中分析海量数据的 IMC 技术，能够迅速获得复杂的分析与交易结果，实时完成业务决策。SAP HANA 在处理逻辑方面采用向量计算的理念，在多核 NUMA 场景下降低功耗，提升多线程性能，并且尽可能多地利用英特尔 x86 CPU 特性和 Cache 高速缓存，以减少内存访问次数，目前技术成熟但成本较高。Apache Spark 是一种大数据并行计算框架，其基于 Hadoop MapReduce 发展而来并扩展了 MapReduce 模型，支持多个工作负载和 Java、Scala、Python 等主流编程语言，突出特色在于基于内存的集群计算技术创新，计算效率相比较 Hadoop MapReduce 在内存中计算快 100 倍、在硬盘数据处理上快 10 倍，现已成为 Apache Software Fundation 中以及关于大数据开源项目中最活跃的项目。

（四）非冯诺依曼架构：量子和类脑成为探索重要方向

量子和类脑是目前颠覆冯诺依曼体系架构的重要探索方向。因基于冯诺依曼体系架构的计算技术升级日渐乏力，通过计算体系架构创新以实现计算能力的快速升级，成为中长期计算技术创新的重要方向。

学术界和产业界持续加大布局力度，抢占未来计算技术领域的新制高点。量子和类脑是目前主要的创新方向，并且与传统冯诺依曼体系都有着本质性差异。量子计算因量子态与原子比特对信息表达方式不同，将彻底改变现有二进制的计算模式；类脑则彻底变革冯氏体系中计算存储和通信之间的逻辑关系。

表3 量子计算、类脑计算与传统计算的对比

类别	经典计算	量子计算	类脑计算
存储	二进制存储器	量子比特（qubit）	人造神经元节点
编码	同一时间仅能存储一个确定的数据	N 个量子比特同时存储 2^N 个数据	同一时间仅能存储一个确定的数据
计算	传统逻辑门实现一次二进制运算	量子门可同时对 2^N 个数进行操作	通过模拟神经元进行分布式存储和分布式计算
优势	技术最为成熟	编码能力指数上升，并行运算，能耗低	大幅度降低计算功耗
瓶颈	冯诺依曼瓶颈制约计算性能，功耗不断上升。	需要提升相干时间和测量精度，算法发展空间受限	目前尚未完全理解人脑构造

多国持续加大科研布局，企业积极探索应用实践。目前美欧日韩等发达国家和地区围绕量子计算和类脑计算均有战略布局和发展计划，通过专项拨款、产学研用等方式加快科研创新。量子计算方面，美国、中国和德国等是全球推动发展的主要力量，在过去十年分别位列全球量子计算论文的前三名，日本、韩国、新加坡等围绕量子信息的研究重点侧重在量子通信，对量子计算只是有所涉猎。类脑计算方面，自上世纪末起美欧等发达国家就以阐明大脑和神经系统机制原理

为目标开展脑科学研究，随后日韩加德英等陆续发布脑科学研究计划，围绕神经形态计算、脑计算等类脑计算领域开展科研布局。在高校及科研机构积极参与计算架构创新研究的同时，科技巨头也在加快推动技术创新、探索应用落地。谷歌、微软、英特尔、IBM 争相布局量子比特制备、量子芯片、量子计算平台等技术生态核心环节，D-Wave、IonQ、Rigetti Computing、1QBit 等初创企业围绕硬件、软件、算法等创新活跃，量子计算技术产业生态正逐渐形成。以此同时，惠普、IBM、英特尔等围绕类脑计算也有相应布局，但整体进展相对缓慢。

量子及类脑等非冯体系计算技术仍处于发展初期，大规模产业化应用尚需时日。量子计算方面，量子比特相比传统计算机比特更强大，50 个量子逻辑比特通用量子计算机性能超过 2016 年全球最强超级计算机“神威•太湖之光”；300 个即可以支持比宇宙中原子数量更多的并行计算。由于量子态叠加、不可复制、退相干等特点，通用量子计算的体系结构不同于经典计算的“冯诺依曼体系”，目前量子计算还处于技术理论验证和原理样机研发阶段，对体系架构、相关软件及算法的研究等均处于起步阶段，预计专用量子计算机将率先得到应用，并在未来相当长的时间内，仍无法完全取代冯氏经典计算机，二者将相辅相成、协同发展。**类脑计算方面，**目前存在两条技术路线，一是基于硅基实现功能类脑，现阶段可实现的计算性能和精度与传统芯片差距较大；二是，基于忆阻器等实现结构类脑，国内外尚无忆阻器原型机，大多处于忆阻器阵列或者芯片研发阶段。总体而言，除脑机接口等交互产品层技术部分取得突破、可实现小范围产业化应用外，基

本原理、硬件实现、软件算法等方面都存在诸多未解决的问题。

我国与全球同期起步，部分技术接近先进水平，但整体与国外差距仍然较大。量子计算方面，中科院作为国内量子计算科研的主力，在 2000 年即开始布局单向光子量子秘钥技术，2005 年实现世界首次光量子 shor 算法，2010 年实现可容错光量子逻辑门，此后在 2015 年不仅采用光量子求解 2×2 线性方程组，也联合阿里创新量子计算实验室。现如今国内量子计算物理实现技术从光学系统向超导系统跨越，已进入超导量子计算技术第一集团。但在量子计算算法、体系结构、编码、材料等方面与国外的差距依然较大。**类脑计算方面，**因目前整体仍处于单点技术突破阶段，我国与全球基本同步。纵观全球，美国在神经形态芯片、核心算法方面暂时领先，我国在基础理论、交互产品方面发展较为突出，是全球首次完成小鼠脑图谱和人类脑图谱绘制的国家，在脑机接口以及智能假肢等产品领域也创新活跃。

四、 近期发展趋势与展望

从第一台计算机 ENIAC 诞生至今，计算技术产业已经走过了七十多年的历程，直接推动了 PC、互联网、云计算和大数据、移动互联网等数轮信息产业发展浪潮。计算技术以基础理论、物理材料、工艺器件等原始创新为开端，先后历经大型机/小型机、PC/服务器、集群/分布式、小型化/低功耗等四大阶段，现阶段虽然非冯诺依曼计算体系创新依然存在，但从近期进展来看仍将以现有软硬件技术协同创新支撑不断升级的应用需求为主，并将随着信息技术产业的进一步升级和与更多传统行业的深度融合，而迎来新一轮发展高潮。

（一） 创新应用是计算技术产业升级的首要驱动力

信息革命加速向万物智能演进，对计算技术提出更高要求。未来人工智能、自动驾驶、VR/AR、5G 等诸多新兴领域对计算需求呈现极速递增态势，并与前五次浪潮有着较大差异。在大型机、小型机和个人电脑时代，对计算的主要需求是采用计算自动化的方式去实现对人工计算的超越和替代；桌面互联网和移动互联网时代，单位时间内实现对多种类计算任务的处理成为最重要需求，如谷歌数据中心中单台服务器每秒需处理 67000 个 RPC 请求，1000 台服务器 7 分钟运行 1000 多个不同应用。万物智能时代对计算的需求则更为复杂：一是数据量的激增对计算性能提出更高要求。人工智能方面，大型人工智能算法所用算力约每 3.5 个月即翻倍；自动驾驶方面，满足 L4 的需求现有计算能力仍需提升 50 倍；VR/AR 方面，现有高端 GPU 芯片的游戏渲染能力难以实现虚拟真实场景的高像素填充率和画面流畅度，图像处理性能至少需要提升 7 倍；5G 方面，若实现 5G 峰值速率超过 20Gbps、端到端延迟降至毫秒级别，现有基带处理能力仍需提升 10 倍。二是不同应用场景对计算能力的需求差异巨大，如物联网应用对计算性能要求适中但对低功耗需求强烈，自动驾驶应用要求低延迟、高并发以保证实时性，VR/AR 则需要高性能和低功耗兼顾。

面向差异化应用需求的优化和加速将是近期先进计算技术产业发展的主要思路。虽然自进入到二十一世纪以来，全球各国政府、科研高校以及巨头企业均不遗余力推动量子计算、类脑计算、新型材料及物理器件的研发创新，但根据目前的发展进度而言，距离大规模产

业化应用仍需较长时间。在此背景之下，至少未来 5-10 年内计算技术产业的创新仍将基于现有硅基冯诺依曼体系展开，摩尔定律的发展滞缓和冯诺依曼瓶颈的日益凸显，使得未来计算资源的稀缺性日益凸显，计算技术产业的升级将更有针对性得围绕具体应用需求得以展开，一方面是通过软硬件技术优化提升整体计算系统的效能表现，另一方面则是摒弃既往通用化发展的思路，面向不同应用计算需求采用专用加速的模式以保证计算对应用创新的支撑作用。

（二） 开放融合是先进计算技术创新的主导模式

先进计算创新是多体系融合的结果。一是数据处理、数据存储和数据交换三大计算单元间的创新融合。硬件方面，计算芯片设计不仅需根据应用需求平衡片上计算和缓存间的配比，还需同步考虑计算芯片间及其与内存等数据存储单元间的 I/O 接口类型；不同计算系统间也需根据应用需求和计算能力的变化，同步升级互联网络连接。软件方面，包括操作系统的文件系统、内存管理以及面向分布式集群的虚拟化、资源管理、任务调度等技术，都将同步耦合升级。**二是**软件技术和硬件技术的创新融合。一方面计算软件将基于对应用需求的拆解进行对硬件能力的适配管理，软件定义的范围和影响力将继续拓展，不仅可实现面向应用的整体系统资源调度和管理，还需针对网络、存储等个性化需求实现软硬解耦和资源灵活配置。另一方面计算硬件将通过与算法和框架等深度融合的专用定制，实现对特定应用需求的支持。目前人工智能专用计算加速芯片创新活跃，未来面向自动驾驶、智能机器人、智能工业装备、智能安防等不同应用领域的需求，将产

生更多样化的计算硬件和计算系统。**三是**围绕计算的信息技术体系化融合。计算技术产业在与通信、传感等信息技术其他领域深度融合的同时，也将深刻影响其他领域的技术创新模式和路径。即将商用的 5G 将全面开启计算和通信融合发展的新纪元，将包括三大融合发展阶段，阶段一是基于计算技术实现网络虚拟化，通过垂直网络切片提高网络资源效率；阶段二是充分发挥终端的计算能力和通信能力，通过终端基站化满足终端灵活组网需求；阶段三是将软件定义网络的功能扩展到终端，实现水平网络切片，使终端通过通信接入实现对固网计算和存储资源的访问调用，以弥补终端计算力的不足，实现各个层面的计算和通信体系化协同发展。

开源开放将成为技术创新的重要手段。随着单一要素对计算技术升级的作用性日益弱化，计算技术创新的难度也将极大提升，与此同时，开源开放发展的理念将使得技术创新的门槛在快速降低。目前借鉴各种开源项目、参与开源社区已成为计算软件技术创新的重要方式，在分布式计算、云计算、大数据和人工智能等领域的重要软件平台都采用开源策略基于开源项目孵化而来，并始终与开源社区紧密捆绑。对硬件开源化创新的探索也不断涌现，因硬件本质上缺乏软件可灵活修改、迭代成本低的特性，截至目前而言硬件领域的开源项目鲜少有成功案例，并且从硬件系统到底层芯片开源的难度逐步加大，RISC-V 是现阶段底层架构及芯片开源的重要尝试，在全球范围内已吸引谷歌、高通、英伟达等超过 100 家单位参与，我国已成立“中国 RISC-V 产业联盟”、“中国开放指令生态(RISC-V)联盟”两大产业组织，以共同

推动开源硬件技术产业体系建设。

（三） 先进计算产业生态进入多元化重构期

多元化将成为近期先进计算产业生态的主要特征。一方面，应用驱动技术创新将加速不同领域的垂直一体化整合。如人工智能领域已形成人工智能应用、算法及开发框架、人工智能芯片为核心环节的垂直体系，包括谷歌、亚马逊等互联网巨头以及英特尔、微软等传统软硬件龙头企业均围绕此方面加大布局力度；自动驾驶领域虽处于起步阶段，但围绕计算芯片、激光雷达等传感硬件、高精地图、核心算法和自动驾驶中控平台等重点环节加速整合。**另一方面，**不同应用领域间的巨大差异将催生更多的细分生态。计算技术演进历程中所形成的wintel、“Android+ARM”等通用统一化生态模式将重构，与人工智能、自动驾驶、VR/AR、物联网等新型创新应用相关的计算生态都将围绕细分应用领域的需求而不断向专用及个性化方向演进。**先进计算产业生态开启多元化重构的同时，也将带来更多的发展机遇。**人工智能及智慧应用驱动计算产业再掀发展新浪潮，预计未来三年依然保持快增态势，五大市场将以超过 50%的增速率先引爆，到 2022 年云端深度学习推理市场规模将超过 20 亿美元、云端深度学习推理市场规模将超过 10 亿美元、人工智能手机市场规模将超过 5 亿美元、智能安防监控市场规模将超过 3 亿美元、智能驾驶汽车市场规模将超过 5 亿美元。应用驱动技术创新、产业升级的同时也将带来计算生态主导者的变化，互联网企业、创新应用企业等在计算生态中的话语权将进一步提升，新兴领域的发展也将为后进入者创造更多发展机遇。

CAICT 中国信通院

中国信息通信研究院

地址：北京市海淀区花园北路 52 号

邮政编码：100191

联系电话：010- 62302881、62304839

传真：010-62304980

网址：www.caict.ac.cn