泛在计算需求服务研究报告

（2022 年）

算网融合产业及标准推进委员会

2022年12月

版权声明

参与编写单位

中讯邮电咨询设计院有限公司

主要撰稿人

黄蔚亭 张桂玉 丁韩宇

前　言

随着数字化时代的不断发展，人工智能、云计算和物联网等技术的兴起，AIGC 和 ChatGPT 等大规模计算模型的应用正引领着各个行业的转型与创新。本白皮书着眼于泛在计算领域的发展趋势和需求。本文将回顾泛在计算产业态势，探讨泛在计算的发展趋势，并详细讨论了泛在计算的需求。同时，我们还将深入探讨泛在算力在典型应用场景中的应用和展望未来的发展方向。

首先，我们将回顾泛在计算产业的态势，探讨过去的发展历程以及取得的成就。这将帮助我们了解泛在计算的发展轨迹，为后续的讨论提供背景和基础。

其次，我们将聚焦泛在计算的发展趋势。在这一部分中，我们将探讨当前泛在计算领域的最新趋势和技术动向，包括算力架构多样性的需求，算网深度融合和确定性的需求，以及算网全要素融合服务和低碳绿色可持续发展的需求等。

最后，我们将详细讨论泛在计算的需求。我们将深入探讨泛在算力在不同领域中的需求，包括异构算力一致性应用场景、云边端算力一体化场景和大规模海量数据调度场景等。这将帮助我们了解不同领域对泛在算力的需求以及如何满足这些需求。

本文旨在为政策制定者、研究人员、行业专业人士和其他对泛在计算领域感兴趣的人士提供一个深入了解泛在计算的平台。我们

希望通过本文的阅读，能够推动泛在计算技术的创新和应用，助力社会和经济的可持续发展。

目　录

图 目 录

一、泛在计算产业态势回顾

多样性算力和融合算力网络是当前云计算和数据中心领域的热门话题，旨在通过集成不同类型的硬件和软件资源来提高计算和网络性能和效率，满足以 AIGC、自动驾驶等新型产业的需求。其中，多样性算力指的是将不同类型的处理器、加速器、存储器和网络设备等资源集成在一起，形成一个具有多种计算和通信能力的计算节点，以适应不同的应用需求。而融合算力网络则是将不同类型的计算节点连接在一起，通过高效的通信网络进行协同计算，以进一步提高系统的计算和通信性能。

在实际应用中，多样性算力和融合算力网络通常需要支持各种不同的算法需求，包括机器学习、深度学习等算法。例如，在深度学习领域，针对图像识别和自然语言处理等任务，通常需要使用大量的浮点计算资源和存储器资源，以满足大规模神经网络的训练和推理需求。而在机器学习领域，常常需要大量的矩阵运算和并行计算能力，以实现数据的快速处理和分析。

为了满足这些算法需求，多样性算力和融合算力网络需要支持不同类型的处理器和加速器，如 CPU、GPU、FPGA 等，以及高速网络和存储设备，以提供更高效的计算和通信性能。此外，还需要有高效的调度和管理系统，以实现任务的自动分配和资源的动态调整，以最大限度地提高计算和网络资源的利用率。

据2022年3月17日，浪潮信息、国际数据公司（IDC）和清华大学联合推出的《2021-2022全球计算力指数评估报告》指出，随着全球数字经济持续稳定增长，数字经济占比预计到2025年有望达到41.5%。 同时，国家计算力指数与GDP的走势呈现出了显著的正相关。15个重点国家的计算力指数平均每提高1点，国家的数字经济和GDP将分别增长3.5‰和1.8‰，预计该趋势在2021年至2025年间将继续保持。从全球算力情况来看，在2020年基础算力规模(FP32)为 313 EFlops，智能算力规模(换算为 FP32) 为 107 EFops，超算算力规模(换算为 FP32) 为 9 EFlops。伴随万物感知、万物互联以及万物智能时代的开启，据 IDC 预测数据，2025 年全球物联网设备数将超过 400 亿台，产生数据量接近 80 ZB，且超过一半的数据需要依赖终端或者边缘的计算能力进行处理。预估未来五年全球算力规模将以超过50%的速度增长，到 2025 年整体规模将达到 3300 EFlops。

图 1

从应用领域来看，首先，泛在算力是智能社会的基石。正如人

均 GDP 是衡量一个国家经济发展程度的重要指标一样，我们采用人均算力指标对主要国家进行了测算：根据目前各国算力发展情况，从低算力国家的 100 GFLOPS/人 到高算力国家的 2,500 GFLOPS/人不等，当前尚处于智能社会的起步阶段，只有当人均算力发展到 10,000 GFLOPS/人 时，才会进入智能社会的发展阶段。（*《中国信通院-中国算力发展指数白皮书》）

其次，构建多元算力生态，能有效促进计算产业发展。智能社会的应用场景多样性和数据类型的多样性对算力提出了多元架构的诉求。要构建繁荣的多元算力生态，既需要底层架构创新，也需要客户、行业伙伴、开发者携手合作，以加速生态的建设，为计算产业发展注入动力。

从行业的角度看，互联网依然是最大的算力需求行业，占整体算力近 50%的份额，以阿里腾讯、百度、字节跳动为代表的互联网巨头对算力的需求更加迫切，同时算力的集中部署也使互联网行业成为先进生产力的代表。政府服务、电信、金融、教育、制造、运输等行业分列二到八位，其中电信、金融行业信息化和数字化起步较早，是我国算力应用较大的传统行业，对算力的应用处于行业领先水平:制造业数字化转型仍处于初期，需要更多规模化、普惠型的公共算力基础设施的支持。

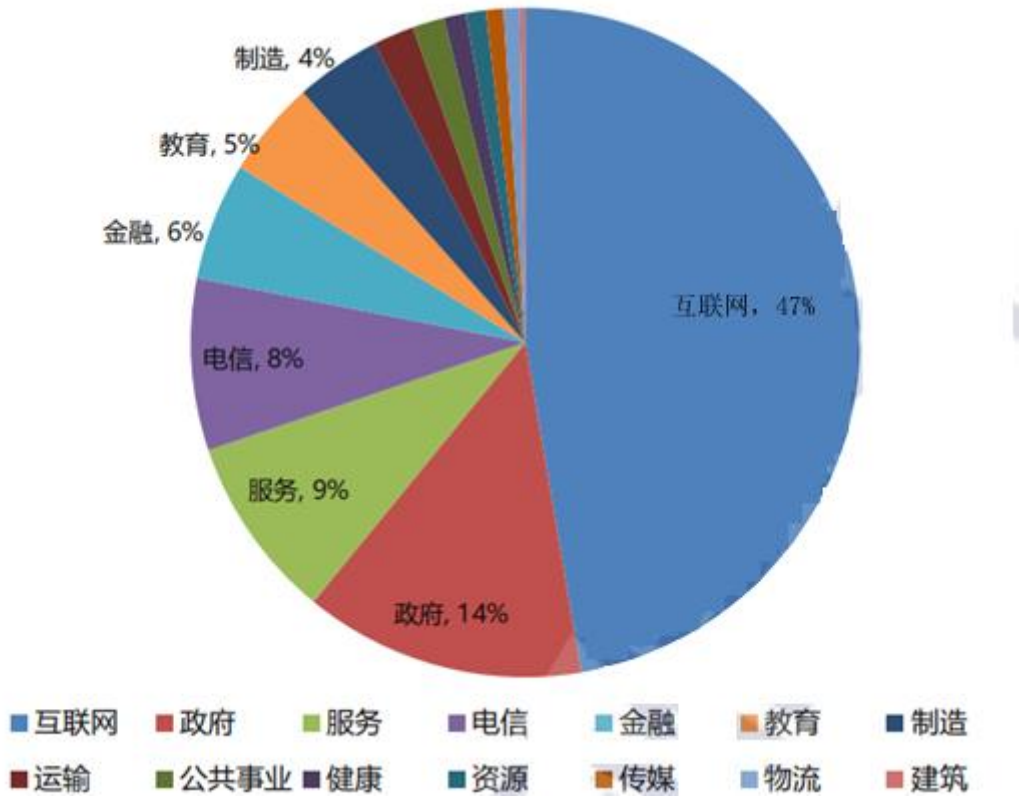

图 2

二、泛在计算发展趋势

AI 大模型的出现是一个划时代的里程碑，人类将进入到一个全新的智能化时代，就像工业革命一样，大模型将会被各行各业广泛应用，带来生产力的巨大提升，并深刻改变我们的生活方式。如同在 2023 年 4 月的阿里云峰会上，阿里巴巴 CEO 张勇所说，所有产品都值得用 AI 重做一遍。未来算网融合和以 GPT 为代表的大模型 AI 能力结合，将会对医疗、工业、教育、电商等行业带来深刻的变革，提高效率、降低成本、提高服务质量，并且为人们带来更

加智能化、高效化的生活方式。AI 大模型通过对产业的数字化转型、产品的 AI 嵌入式改造等方式，成未来数字经济的智能大脑。而这个数字经济的底座是算力和网络，要真正实现“全国算力一张网”、“像用水用电一样随需随用算力”，还需要将算力和网络资源整合为一个整体，形成一个高效、可扩展、灵活的泛在算力网络，这都要求新的 ICT 格局向着泛在互联与泛在计算统一调度统一控制的融合方向演进。

算力分布化和异构平台专业化的发展趋势越来越明显。这种趋势驱动了网络功能虚拟化（NFV）和软件定义网络（SDN）的发展，以加速云化和分布式计算。同时，由于不同应用对算力和网络的要求不同，因此异构平台和分布式计算的需求也随之增加。

服务 AI 化是一个重要的趋势，许多企业都在致力于将其服务转化为 AI 驱动的服务，以提高生产效率和用户体验。这些 AI 服务需要更强大的算力来支持它们的运行和训练。

技术趋势方面，网络带宽需求将会超过摩尔定律的发展速度，因此更加追求有效算力。有效算力推动系统整体的架构设计、集成优化，软硬协同和软件层的技术优势。与传统的数据中心算力规模衡量指标，如柜数、硬件规格等相比，有效算力有利于提高大数据中心应用场景的设计与优化。

另外，随着 5G 和物联网的发展，将会有更多的设备和传感器连接到网络中，这将进一步推动分布式算力的发展。分布式算力能

够在边缘设备和云端之间平衡计算负载，提高应用的响应速度和效率。

最后，从硬件角度来看，异构计算架构和专用硬件（如 GPU、TPU 等）将会成为发展趋势。这些硬件将能够提供更高效的算力和更低的能耗，以满足不同应用的需求。

三、泛在计算的需求

数字经济的发展，产业与 AI 结合的应用，对于算力和网络都提出了很高的要求，由此催生出了“泛在算力”这种新的技术形态和运营模式。泛在算力网络的未来发展路径，将是算和网系统化融合、统一服务的过程，是从位置与运营层面的融合，到控制面的融合，最终将会实现技术、协议和设备形态的融合。数字经济的发展将推动海量数据产生，数据处理需要云边端协同的强大算力和广泛覆盖的网络连接。多样性算力、算网融合等将成为重要趋势，不同的应用需要不同的算力支持，从 CPU、GPU、TPU 等硬件架构到分布式计算、高速互联等技术手段，都需要在不同的场景中进行选择和应用：

1. 泛在算力架构多样性的需求

《数字中国建设整体布局规划》中除了强调“系统优化算力基础设施布局”、“加强传统基础设施数字化、智能化改造” 等重大部署

外，还包含了“构建国家数据资源体系，健全各级数据统筹管理机构”等对于数据资源体系建设的纲领性要求，全社会数字化转型带来多样化的海量数据处理需求，对处理效率提出了更高的要求，算力呈现出内核架构多样化、应用特异化的趋势。传统的以 CPU 为中心的计算架构难以高效应对复杂的数话处理场景，以数据为中心的新型多样化计算架构正在迅速兴起，以 GPU、FPGA 芯片为代表的异构算力逐渐成为主流，以 DPU、NPU 代表的软硬件深度融合、定制化一体化计算架构迅猛发展，突破传统诺伊曼架构的近存计算、存算一体等存算融合新计算架构不断出现。这些架构变革以数据处理的高效性为目标，通过数据流驱动计算，对底层数据按需就近处理，极大提升数据处理效率，从而由量变到质变，典型的如 chatGPT 等大模型的兴起，OpenAI 等企业并不是在理论上实现突破，而是在工程技术上的突破实现了超大规模不间断的算力输出。此外，面向多样性的新型计算架构，跨架构的开放编译平台已成趋势，通过屏蔽底层硬件架构差异，构筑开发环境友好、性能高效的算力服务。

2. 算网深度融合与确定性的需求

传统的算力和网络相对独立，二者仅为简单的连接与承载关系。网络的发展让算力更易泛在扩展，让数据更易流动，用户更便捷使用。算力要发挥极致性能，就势必要求网络技术变革创新，从 TCP/IP 体系的“尽力而为”到 RDMA 等体系的“无损网络”，已经在数据中心

内有了大规模的实践，也为行业带来了巨大的变革。未来，进一步驱动网络感知算力的位置，实现就近分流，算网融合的一体化平台服务已成趋势。通过网络连接泛在算力，可突破单点算力的性能极限，发挥算力的集群优势，提升算力的规模效能，通过对算网资源的全局智能调度和优化，可有效促进算力的"流动"，满足业务对算力随需使用的需求。同时，伴随着行业应用对网络在端到端质量方面的极致要求，网络需从尽力而为向端到端确定性保障演进，网络协议也需创新发展。例如，算网将在协议和形态层面进深度的交互，网络将深度感知算力，通过在网络协议中引入算力信息，将应用请求沿最优路径调度至算力节点；随着产业数字化转型进程加速，网络正从消费互联网向工业互联网演进，以超低时延和确定性为特征的网络正从辅助生产逐步嵌入到核心生产环节，成为产业数字化刚需。确定性要素从带宽向时延、抖动、丢包等多要素转变，满足算网连接业务多维质量需求。确定性范围逐步从局域走向广域，通过 FlexE 切片和 SRv6/G-SRv6 实现端到端确定性保障；确定性粒度逐步从粗粒度向精细化转变，通过小颗粒切片、应用感知等技术满足差异化服务体验，这些变化，将推动 V2X、云手机、云游戏、VR 等超低时延类新型业务创新突破。

3. 算网全要素融合服务的需求

算力将成为多技术融合、多领域协同的重要载体。全行业数字化转型的加速对算网一体化基础设施和融合人工智能、大数据等多要素融合服务能力的要求日益提升。当前算网各自编排、分域管理，难以提供算网融合的产品、服务和端到端的质量保障。算网大脑通过算网数据感知获取全域实时动态数据，结合算网智能化、多要素融合编排实现要素能力的一体供给和智能匹配，横向全面融合网、云、数、智、安、边、端、链（ABCDNETS）多种能力要素（*《中国移动算力网络白皮书》），纵向深度贯穿应用、平台到底层资源，进而为新型信息基础设施对外提供一体化服务提供能力支撑。进一步提升算力服务的智能化水平、可信交易能力，推动算力服务向纵深发展。

4. 低碳绿色可持续的泛在算力需求

2021 年底我国在用数据中心规模达到 520 万标准机架，算力达到 120EFLOPS，未来仍将保持每年 10%以上的速度增长。同时，数据中心能耗快速增长，2019 年数据中心年耗电量占全国总用电量的 2.4%左右，预计到 2025 年数据中心年耗电量将达到全社会用电量的 3.6%。为实现可持续发展，实现国家双碳目标，算网基础设施的建设和发展要以绿色低碳贯穿始终。积极推动“东数西算”的发展，跨地域、跨层次、跨架构的海量数据调度，以 OXC、400G/800G、OSU、SPN、50G PON、FTTR、硅光、新型光纤等为代表的新技术

推动光传送网向基于光电联动的大带宽、扁平化、低时延的智慧全光网演进。通过大容量全光高速互联、灵活光电联动和智能全光调度，围绕算力构建高容量、高可靠、确定性低时延、绿色节能的光网络，以光筑底，支撑东数西算新型算力服务的创新发展。

随着芯片、服务器能耗控制技术的日趋成熟，模块化、工业化等新型建造技术的规模推广，高效、节能制冷技术的推陈出新，融合、低碳能源技术的广泛应用，辅以精准的碳评估与管理手段，从而形成从芯片、主设备到基础配套设施全生命周期、绿色低碳的数据中心建设与运营体系，并以数据中心为核心，结合数据中心间的网络连接，充分利用智能化手段实现网络和算力的跨地域、跨领域的节能调度协同，形成端到端的一体化节能体系，最终实现算力网络的创新节能、智慧洁能、绿色赋能，助力数字社会可持续发展。

5. 全程可信、共享的泛在算力需求

随着信息技术的发展，算力和数据的流动性持续增强，新型泛在算力服务需要多方可信的算力交易和安全可控的数据流通，这个行业将重塑信息服务产业价值链分配体系。通过算力交易创新模式广泛吸纳社会算力，基于区块链的多方算力可信共享将推动算力的供给侧改革，使算力的使用成本进一步降低，实现算力普惠，同时提升社会空闲算力使用效率，最大化发挥算力的价值。数据流通服务可实现跨行业、跨主体的数据共享和开放，通过联邦学习、隐私计

算等技术为多方协作模式下的数据价值挖掘提供可靠保障，推动数据合法、高效利用，助力数据产业应用升级，打造数据应用服务新范式。同时，从云计算到边缘计算再到分布式云，算力的泛在化引入了更多的安全风险点，更加开放的网络架构和更大范围的数据流动导致不确定性安全威胁增加，传统以安全防护为主的“外挂式”或“补丁式”安全建设模式无法应对上述安全问题，以安全能力内生、安全可信为基础的新安全理念应运而生。在资源高度协同、网络灵活开放、数据高速流通的算网环境中，充分应对动态变化的安全需求，引入安全编排、隐私计算、全程可信等技术，提升安全风险自动发现、自动防御的能力。

四、泛在算力的典型应用场景

泛在算力是指所有能提供计算能力的设备都可以被整合利用的概念。在大带宽方面，泛在算力的应用可以使计算任务可以更快速地传输和处理，实现更高效的数据处理；在确定性时延方面，泛在算力的应用可以确保数据在传输和处理过程中不会出现延迟，实现更高效的数据处理和应用场景；在高算力方面，泛在算力的应用可以提供更强大的计算能力，从而实现更复杂和更高效的数据处理和应用场景；在移动化方面，泛在算力的应用可以让移动设备拥有更强大的计算能力，实现更多样化的移动应用场景；在可信方面，泛

在算力的应用可以提供更高的安全性和可靠性，保证数据和应用的安全性和可靠性；在分布式方面，泛在算力的应用可以让多个设备协同工作，实现更大规模和更高效的数据处理和应用场景。

1. 异构算力一致性应用场景

从 2017 年至今，互联网上的应用发展形态呈现从图文、点播逐步走向更加实时的直播、实时音视频、AR/VR。与此同时，应用对网络需求以及算力、算法形态也随之不断革新。随着技术的不断演进和创新，例如基于大数据的深度学习、智能物联网和 5G 等新兴技术，泛在算力网络将会更加智能化和自适应，支持更加丰富和多样化的应用场景。

按 Sandvine 统计，2022H1 除视频和游戏类之外的数据占互联网流量近 30%。排除掉互联网用户增加、视频分辨率提升、用户在网时长增长等影响因素，如果生成式 AI 能够将基于文字、图片和数据文件的信息传递模式“升级”到视频类，则由于媒介不同必将使得网络流量总量提升，据此推论，视频流量占比会超过 80%。

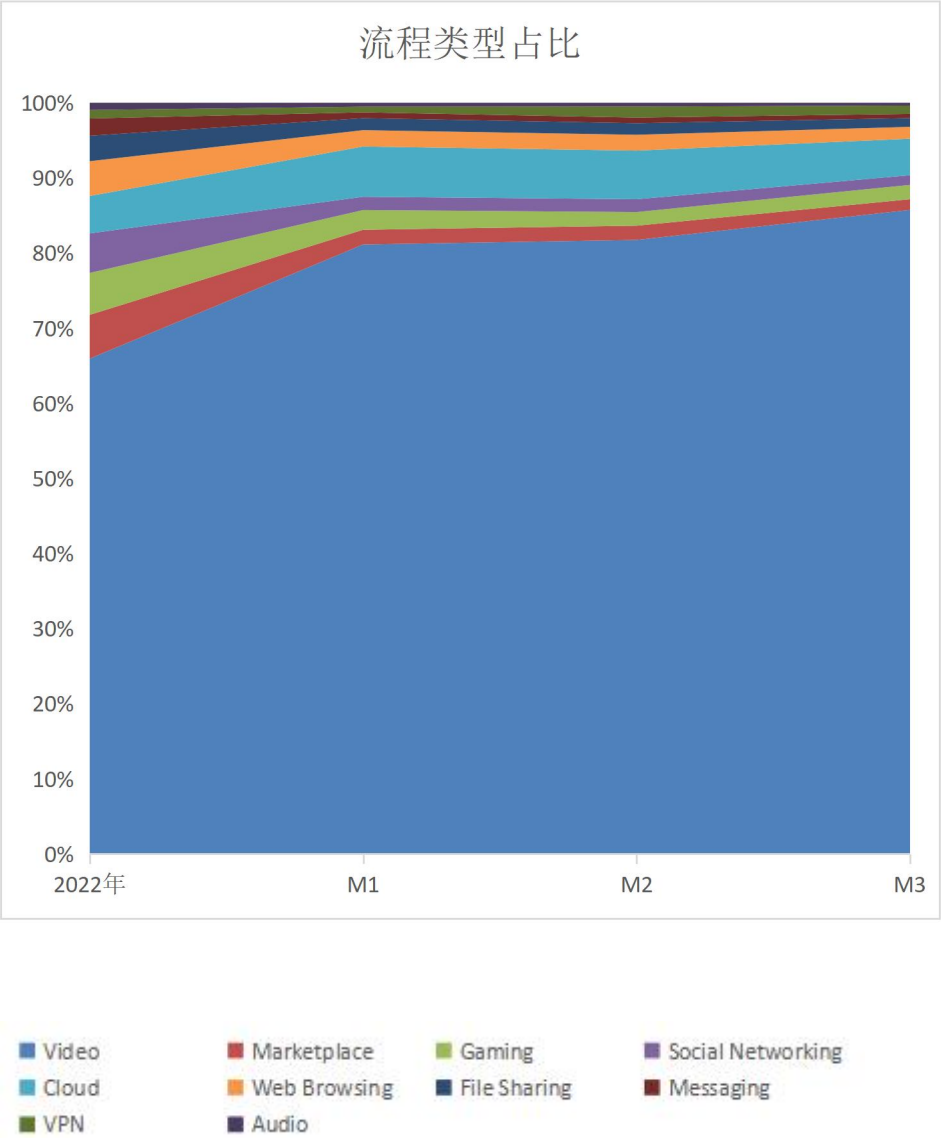

图 3

移动互联网的应用，从图文类到视频类，再到 AR/VR 和 AI 生成的视频，对应算力的需求是从物理机为代表的内容分发网络，到虚拟机、容器为代表的虚拟化平台，再到 CPU+GPU+DPU 组成异构平台所提供渲染、推理计算一体化平台。这样架构的迭代，在应用到体验上，确保用户的一致性。

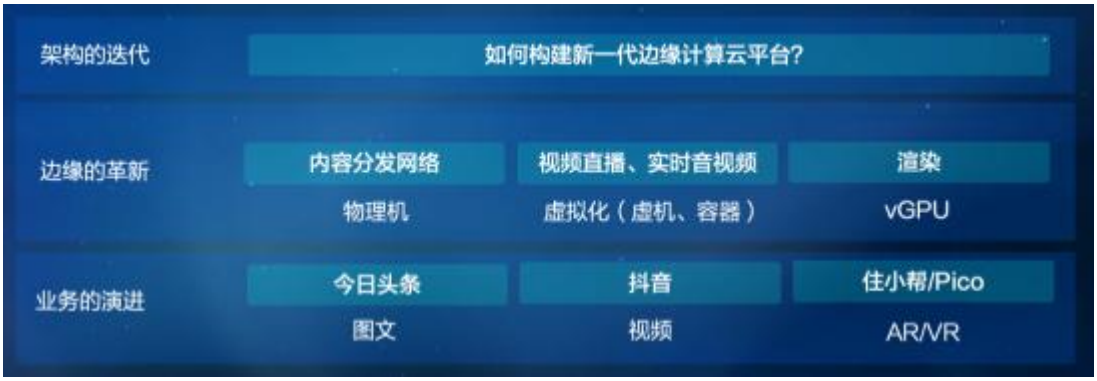

图 4

随着人工智能应用的蓬勃兴起和大规模发展，对智能算力和超算算力的需求与日俱增，AI 算力的核心地位日益凸显。异构计算不再以渲染、训练等离线场景为主，逐步发展以推理为主的准实时的算力服务应用。

随着 AI 应用在各个行业的不断深入，数据规模、算法复杂度、企业业务场景多样性都呈几何倍数增加，常见的异构计算平台分类与应用的关系如下：

图 5

除了常见的 CPU+GPU 组合，FPGA 在最近两三年已经在图片转

码、视频转码与人工智能推理等领域得到充分验证。FPGA 具备高性能、低延迟、高扩展性、高能效比、高灵活性等特点。由于 FPGA 摆脱了冯诺依曼结构的限制，可以针对特定应用算法定制硬件架构，具备 ASIC 高性能的特质，可支持更高的数据吞吐，FPGA 不但可以实现数据并行，还可以实现流水线并行，可达到微秒级乃至更低的延时，这也是其与同样高算力的 GPU 相比最明显的一个竞争优势。基于 FPGA 的可编程性和丰富的 IO 管脚，其可以在数据中心里面扮演多面手，除了计算，也可支持和适应存储、网络等方面算法的演进发展要求。更先进的半导体工艺，加之优化的算法，令 FPGA 可在特定应用中实现比 CPU、GPU 更佳的性能功耗比。另外在灵活性上，FPGA 与 ASIC 相比最引以为豪的强项就是可编程性，而且 FPGA 适于与 CPU 等计算架构结合形成 CPU+FPGA 异构计算平台，让最合适的架构去做其最擅长计算加速，以实现系统最优。这些能力让 FPGA 在处理 JPEG 到 WebP 和 JPEG 到 HEIF 图片的转码实践中，FPGA 处理延时相比 CPU 降低一到两个数量级，处理性能则是 CPU 的 6 倍（Intel Skylake 96HT），TCO 节省至少 30%。同时，在视频解码方面，新视频编解码标准 H.265 在保持画质不变的前提下，码率只有 H.264 的一半，但是编码所需要的算力是 H.264 编码所需算力额的 3 到 5 倍，CPU 对于这类 1080p 以上清晰度的内容编码已经力不从心。FPGA 独有的流水+并行计算优势在高清视频转码中表现出了巨大的优势，720p 的视频流采用 H.265，相比 H.264 在提升

了画质的同时，节省了至少 20%的带宽，大幅降低了带宽成本。

针对 CNN 及视觉类算法，NPU 可以在图像搜索、图像识别、视频内容识别、自然语言识别等场景做到高效低能耗，目前已经广泛应用在城市大脑、视频图片合规审查应用。

在智慧城市的应用中，交通信号机系统使用 NPU 服务器处理车辆检测、跟踪、车辆品牌识别、车牌识别等算法模型，单张 NPU 卡全链路能够支持 100 路实时视频的分析和特征结构化数据的提取，在算法精度达到同样效果的情况下，端到端性价比是 GPU 的 5 倍。

在短视频、电商图片与视频审核上，基于 NPU 芯片上，在算法精度达到同样效果的情况下，端到端性价比是 GPU 的 4 倍。

同时，在资源管理方面，异构算力资源的建设也需要考虑高效的资源管理策略。这包括如何对不同特点的算力资源进行统一的管理和调度，如何平衡各个节点的负载，以及如何进行资源的动态调配，实现更细粒度的资源管理，提高计算资源利用率，达到降本增效的目的。算力和算法之间的关系是相互促进和相互制约的。算力提供了算法执行的基础，而优秀的算法则可以更加高效地利用算力资源。算法就是支持整个智能世界的灵魂，算力是承载算法的物理基础。业务层网络由业务处理节点和节点之间的连接组成，未来网络要从信息传输为核心的信息基础设施，向融合感知、传输、存储、计算、处理为一体的智能化信息基础设施发生转变(来源：中国人工智能 2030 战略规划)，这对业务处理节点的功能、节点之间的连接技术

都提出了新的要求。

从应用需求出发，未来网络架构，需要能够支持不同的计算功能，根据不同的业务需求、网络状况，可以在离客户端的不同距离实时地实例化，同时基于云原生架构，通过以更加松耦合的理念，在应用资源上融合不同的异构算力，在应用管理上支持更加灵活的编排调度，在应用服务上支持云边端一致的用户体验。同时，通过云原生可以打通从设计、开发、集成、测试、发布、部署、运维、监控的产品全生命周期链路。实现用户体验最优、计算资源利用率最优、网络效率最优。

2. 云边端算力一体化场景

全球数据总量仍在持续增长，预计 2020 年达到 47 ZB，2025 年达到 163 ZB，年复合增长率为 20%，其中绝大多数来自亚太（约 40%）和北美（约 25%）、欧（约 15%）地区。全球数据中心安装服务器数量 2020 年将达到 6200 万台，年增长约 4%；智能终端（包含手机/M2M/PC 等）年复合增长约 10%；由于工艺的约束，单芯片的算力在 5nm 之后将接近顶峰，传统集约化的数据中心算力和智能终端的算力可增长的空间也面临极大挑战。要支持数据持续增长的机器智能时代，只有终端+数据中心两级处理无法满足要求，算力必然会从云和端向网络边缘进行扩散。数据处理会出现三级架构：终端、边缘和数据中心，边缘处理能力未来几年将高速增长，尤其

是随着 5G 网络的全面建设，其大带宽和低时延的特征，将加速算力需求从端、云向边缘的扩散。

算力按照应用场景有不同的衡量单位，用于比特币的每秒哈希运算次数(H/S)，用于 AI 和图形处理的每秒浮点运算次数（FLOP/S），智能社会对算力的诉求主要是浮点运算能力，专用 AI 芯片如华为昇腾 910 采用 7nm 工艺，半精度 FP16 算力达 256 TFLOPS，低功耗的 12nm 芯片昇腾 310 半精度 FP16 算力也达到了 8 TFLOPS。过去 5 年，随着深度学习算法的演进，AI 训练对算力的需求增加了 30 万倍，一些互联网厂家已经将算力作为服务提供给客户，从 1 FP32 TFLOPS 或 8 FP16 TFLOPS 到 4 FP32 TFLOPS 或 32 FP16 TFLOPS 的 AI 推理加速服务，简单的语音语义识别或单流视频分析 8 FP16 TFLOPS 即可满足，复杂的推荐引擎或者风险检测则需要 32 FP16 TFLOPS。

表 1 推理不同应用对算力的需求

AI 推理场景	算力分类	32 位浮点运算	16 位浮点运算	内存
语音语义或者单流视频	中等	1 TELOPS	8 TELOPS	1 GB
多流视频	大型	2 TELOPS	16 TELOPS	2 GB
推荐引擎或者风险预测	超大型	4 TELOPS	32 TELOPS	4 GB

对应算力分类，可以将用户到云中心之间所有的算力层分成三类：

现场端、近场边、云中心三层。

- 首先，“现场端”，主要位于用户现场或用户自己的机房，覆盖 1 到 5ms 时延范围，可以将云中心训练好的模型算法和能力下沉到用户的现场侧，满足超低时延的计算和网络能力。现场边缘主要应用于 IoT、边缘时序数据等实时性业务的典型场景。
- 其次，“近场边”，主要位于全国二三四线城市或城区节点，覆盖 5 到 20ms 时延范围。目前近场边缘主要在 CDN、视频直播、实时音视频、视频监控和图像处理等常见业务场景落地。
- 最后，“云中心”，位于区域中心城市、提供多线及 BGP 汇聚节点，覆盖 20-40ms 时延范围，可以跟中心云实现高效连接，为“现场边缘和近场边缘”提供汇聚能力等。目前云边缘在 CDN 合并回源、视频直播的 L2 层转发、离线渲染业务、数据并发处理业务等场景有广泛应用。

在云原生的架构下，终端应用可以实现更轻量的应用形态和更友好的硬件支持，边缘计算则提供更实时的服务响应和更精准的用户覆盖，而中心云计算则能够实现更高效的数据聚合和更敏捷的业务架构，最终达到云、边、端三位一体，协同一致的目标：

2.1 视频直播场景

在视频直播场景中，边缘节点可以帮助业务实现直播流的就近分发和就近访问，确保直播的低时延，降低中心带宽压力。同时，边缘节点能够支持实时弹幕的边缘分发，在靠近观众侧实现高效拉流，提升主播、观众双向的直播体验。

基于高质量的画面诉求，低时延的转码也是直播场景中的关键因素，丰富、高性能的边缘算力能够满足直播中不同业务的多样化算力资源需求。此外，边缘计算云平台具备的 VF 直通功能可以减少虚拟化对网卡转发能力的损耗，IPv4/ IPv6 双栈、负载均衡、镜像预热等能满足直播业务所需的主要功能和快速全域部署的能力，真正为用户提供高清、流畅的直播互动和观看体验。

2.2 实时音视频场景

随着视频会议、在线教育等场景的普及，端到端之间实时互动的要求要越来越高。实时音视频可以借助边缘节点实现业务的就近接入，保证节点间低时延互联互通，提供高速稳定的实时音视频通信优质链路。同时，边缘算力的弹性扩容能力能保障业务量突增时，视频会议中长会话的通信质量，而边缘计算 GPU 实例还可以满足实时音视频中的渲染需求。在功能上，高性能负载均衡可以支持实时音视频在边缘节点内高效东西转发，打通东西向流量。此外，多线、IPv4/ IPv6 双栈等也为实时音视频提供完整的能力保障，满足多人连麦、多人视频会议的低时延需求。

2.3 边缘渲染场景

在边缘渲染场景中，如常见的直播特效、家装应用涉及的 3D 特效和 VR 看房等，在内容制作环节往往有大量的工程数据需要处理。边缘计算可以基于设计师所在地理位置就近提供服务，缩短工程数据传输距离，有效降低网络时延，提高业务渲染的实时性。

同时，通过全域节点的边缘算力资源和智能调度，能满足关键渲染任务的灵活切片，实现多节点并行渲染，提升渲染效率。

2.4 云游戏

云游戏场景中，用户对时延更加敏感。区别于端游、页游、手游和主机游戏，云游戏的游戏资源、运行、渲染都需要在云端完成，相当于用户在云端玩游戏。

云游戏业务依托全域覆盖的边缘异构算力，基于用户地理位置的亲和性，通过边缘智能就近调度，实现游戏指令毫秒级交互。同时，结合高密度的 ARM 集群、GPU 算力、弹性扩缩容、资源隔离等功能，支持多个云游戏实例并发运行，为终端用户提供无设备限制、稳定、高品质、超低时延的游戏体验。

2.5 边缘函数

常见的边缘函数场景如：浏览器性能优化、页面个性化内容的生成、A/B 测试和边缘鉴权的处理等，为了降低程序部署和批量发布

的时间、成本以及用户的编程门槛，边缘函数支持 JavaScript 调用浏览器运行时 API，可快速编写代码或调用通用模板，实现一键式全球下发部署。

同时，边缘函数能够快速响应客户的 HTTP 请求，就近调度到边缘节点执行，整个启动时间可控制在 3-5ms；还能够配合 CDN 实现如：鉴权、边缘定制应用等服务；以及源站拨测的探针，如 A/B 测试等；实现业务的快速分析和决策。

另外，当节点的客户端请求数量激增时，平台还支持将请求有序调度至周边充足的计算节点处理，实现快速、高效的扩容和调度的自动化管理，并通过提供更细粒度的弹性资源，实现多租户函数工作流环境隔离。

2.6 V2X 广泛丰富的计算需求

所谓 V2X，意为 vehicle to everything，即车对外界的信息交换。V2X 的含义不仅仅局限于车联网，还包括借助新一代信息和通信技术，实现车与车、车与路、车与人、车与云端的全方位网络连接。通过收集车辆、道路、环境等信息，使汽车和其他对象智能协同配合，提升汽车智能化水平和自动驾驶能力，构建汽车和交通服务新业态，为用户提供智能、舒适、安全、节能、高效的综合服务。根据亿欧数据，L1 级别算力需求小于 1TOPS，L2 级别算力需求 2TOPS，L3 级别算力需求 30TOPS，L4 级别算力需求 300TOPS，

L5 级别算力需求 4000+TOPS。从 V2X 的发展来看，域包括以下细分领域：

- 智能交通管理：通过利用算力网络分析交通流量、车辆速度和道路拥堵等信息，为交通管理部门提供决策支持，例如调整交通信号灯、限制车速和重定向交通流。
- 自动驾驶车辆：无人驾驶车辆需要实时处理传感器数据、生成驾驶决策，并执行控制操作。算力网络和算力服务可以为自动驾驶车辆提供强大的计算能力，支持自动驾驶决策和实时控制。
- 高清地图构建：无人驾驶车辆需要精确的高清地图来导航和定位。算力服务可以利用多源数据构建高精度的数字地图，支持自动驾驶车辆的定位和路径规划。
- 车辆安全性分析：通过分析无人驾驶车辆的传感器数据，算力服务可以识别潜在的安全风险和故障，并提供预测性维护和修复建议。
- 交互式驾驶体验：算力服务可以为车内娱乐和信息娱乐系统提供强大的计算能力，支持多媒体内容的实时流媒体和个性化推荐。

而以上的场景中，包含了供车辆终端业务实时计算转发、离线计算能力，包括数据解析、实时计算、消息转发推送、离线计算等算网融合的需求。同时作为支持平台的智能交通管理模块上，需要应

对高并发、低延迟的要求，依据现有可参考的国际和国内标准，以基本的 5 大类信息 BSM、SPAT、MAP、RSIRSM 进行估算，根据平台服务范围进行分析，终端接入平台的并发量预计如下:

- 区县范围: V2X 平台需提供支持每秒百万条数据并发接入的能力
- 城市范围: V2X 平台需提供支持每秒千万条数据并发接入的能力
- 省/全国范围: V2X 平台需提供支持每秒上亿条数据并发接入的能力。

综上，V2X 的计算架构逐步走向区域化集中计算方式，集成化的设计可以降低算力冗余要求的同时大幅降低整车线束长度，有效降低成本。当域集中之后，智能化功能升级将从增加传感器数量转为增加算力、算法模型和数据训练，因此对自动驾驶 AI 芯片算力要求将越来越高，每提升一个级别，算力需求增加 10 倍以上。根据亿欧数据，中国自动驾驶 AI 芯片市场规模 2021 年为 25.1 亿元，预计到 2025 年将达到 109.9 亿元，CAGR 为 44.7%。

3. 大规模海量数据调度场景

早在 2016 年 10 月，习近平总书记在中央政治局第 36 次集体学习时便提出“以数据集中和共享为途径，建设全国一体化的国家大数据中心”的战略方针。近两年来，国家发改委等四部委围绕全国一体

化大数据中心体系印发了多份重磅文件，提出要大力发展数据中心集群，加快实施国家“东数西算”工程，开展数据中心与网络、云计算、大数据之间的协同建设，促进解决东西部算力供需失衡问题。《数字中国建设整体布局规划》的发布，首次从国家数字化层面出发制定的一个系统性的规划，具体的提出要系统优化算力基础设施布局，促进东西部算力高效互补和协同联动，引导通用数据中心、超算中心、智能计算中心、边缘数据中心等合理梯次布局。《规划》的发布，恰逢以 OpenAI 为代表的 GPT 大模型在以天为单位快速迭代的时刻，随后国产 GPT 大模型雨后春笋般推出，对算力的要求逐步增加。

在公共大模型场景下，据研判，在 2023 年到 2024 年之间，GPT 将以 2C2H 消费型为主，即低频文本非即时交互为主，时延不敏感。这个阶段大模型训练和推理均集中部署，例如在东部枢纽节点等，此时 GPT 对网络的大带宽需求，并不是特别迫切。但随着生产型流量超过消费型，预计 2027 年前后，高频富媒体即时交互为主，时延敏感且带宽需求量较大，生成式 AI 的模态发展对网络带宽需求会超出摩尔定律，算力布局将逐步形成以训练集中部署在西部枢纽节点，推理分布式部署的架构。此时行业呈现出户数量继续持续渗透，集中训练后的模型需要同步到分布式推理端，骨干网流量增加、用户数据隔离、城域段流量潮汐明显等态势。

在垂直领域小模型场景下，对于大中型企业出于数据安全考虑

无法使用公共云上的 ChatGPT 服务，因此最常见的方案是部署一套私有化的垂直领域大模型，并使用公司内部积累的数据进行训练，最终实现对本行业的推理应用。但这个过程中，往往耗资巨大，仅用于训练的 GPU 投资就在千万级以上，训练过程的电力成本和调优过程中的人力成本也是个不可忽视的巨大成本。由于训练过程往往只持续一到两个月，所以有些公有云逐步发展出专有可用区租赁的方案，即不同的客户按月租用一整个可用区，在租用过程中，用户的数据中心与公有云可用区进行 IB 网络互联，此方案基于最远可达 10km 的长距 ROCEv2/IB 网络来实现。

基于大模型训练，基于专业数据调优的垂直领域小模型，目前已经在很多领域发挥了重要作用。融合算力网络和算力服务可以为医疗领域提供高性能计算和存储能力，帮助医学研究和诊断变得更加准确和高效。在 2020 年疫情肆虐初期，阿里云就通过机器学习 AI 能力，来协助医生查看 CT 影像，并以 73%的准确率胜过了 94%的专业医生，且在用时上远远少于人类。常见的生物医学方面，算力加持下的 AI 能力还应用在：

- 医学图像分析：利用算力服务提供的高性能计算能力，可以对医学图像进行深度学习算法分析，例如 CT、MRI 和 X 光图像等，以协助医生快速准确地进行疾病诊断。
- 基因组学研究：基因组学是一个数据密集型领域，涉及到海量的基因数据分析。通过利用算力服务提供的高性能计算能

力，可以加速基因组数据的分析和研究，进而促进研究成果的发展。

- 医疗数据管理：医疗数据管理涉及到大量的数据存储和处理。通过利用算力服务提供的高性能存储和计算能力，可以实现大规模医疗数据的存储和分析，支持医院进行数据管理和应用。
- 云诊断：通过将医疗设备连接到算力网络中，可以实现云端医学影像分析和诊断，即通过医疗设备采集的数据传输至云端进行分析和诊断。这种方式可以提高医学影像的准确性和快速性，以及减轻医生的工作负担。

五、总结与展望

算力需求指数级增长，对数据中心网络提出了更高的要求，云化成为了必然之路。人工智能算法的复杂度和计算量不断增加，需要更高效、更强大的算力支持。这就要求算力网络必须具备更高的处理速度、更低的延迟和更大的计算能力，才能满足不断增长的计算需求。例如，GPT-3 的大模型的所需要训练 355 个 GPU-年。在 GPT 成果以“天”为单位迭代的过程中，科技公司用于训练 AI 大模型的时间为 1 个月，因此其需要训练 AI 大模型的 AI 加速卡的数量为 4260 个，这是一个相当巨大的算力需求，基于对数据中心的统筹发展，

国家发改委在《关于加快构建全国一体化大数据中心协同创新体系的指导意见》中提出“布局大数据中心国家枢纽节点，形成全国算力枢纽体系”的具体要求。其中特别指出，构建一体化算力服务体系，加快建立完善云资源接入和一体化调度机制，以云服务方式提供算力资源，降低算力使用成本和门槛。支持建设高水平云服务平台，进一步提升资源调度能力。同时，还要优化算力资源需求结构。以应用为导向，充分发挥云集约调度优势，引导各行业合理使用算力资源，提升基础设施利用效能。对于需后台加工存储、对网络时延要求不高的业务，支持向能源丰富、气候适宜地区的数据中心集群调度；对于面向高频次业务调用、对网络时延要求极高的业务，支持向城市级高性能、边缘数据中心调度；对于其它算力需求，支持向本区域内数据中心集群调度。在这个基础上，可以通过以下三点实现：通过引入云原生技术，实现业务逻辑和底层资源的解耦，释放开发者的活力；面向服务的容器编排调度能力，实现服务编排面向算网资源的能力开放；社会中多产权主体可提供多种异构算力，实现对泛在计算能力的统一纳管。

随着应用场景的不断扩展，算力资源需要支持分布化。实现更多的设备类型和接入方式，包括智能终端、物联网设备、车联网设备等。这就需要算力网络具备更高的灵活性和可扩展性，能够支持各种类型的设备接入和协同计算。随着以 V2X、无人机等高速移动大带宽持续连接的需求不断增加，出现了以移动设备和物联网设备为

主的端侧计算。虽然网络化的计算已经有效补充了单设备无法满足的大部分算力需求，但是由于不同类型网络带宽及时延限制，仍然有部分计算任务需要更高的计算能力。因此，未来将形成“云、边、端”多级计算部署方案，即云侧负责大体量复杂的计算，边缘侧负责简单的计算和执行，终端侧负责感知交互的泛在计算模式。新的 ICT 格局将向着泛在联接与泛在计算紧密结合的方向演进。

以 IPv6 提供基础联接，增加 SRv6、随流检测等创新技术支持网络感知应用能力，真正实现“网络即服务”。以上的两个趋势对网络提出了要具备智能化、一体化的云网服务体系，满足客户一键式订购云网产品的需求，同时能够自动匹配网络资源，快速开通，业务灵活调整，面向最终客户提供云网一体化的产品与服务。客户可以像在电商平台采购一样任意选择产品组合，从签约到履约实现在线自助，快速开通，流程可视，实现业务开通自助化，同时还可以协助运营商和企业减少人力成本。此外，在企业内部，针对重点应用需要保障与服务，还需要进行 SLA 分级，对应确定性网络。新技术如 SRv6 和 APN6 等可以提供这些所需的支持，从而实现数字化转型 后，应用的云原生化和智能化，以及网络的智能化和一体化。这样的新型网络支持将更好地适应数字化时代的业务需求，并为企业提供更加灵活、高效、可靠的服务。

泛在的算力网络融合的核心思想是通过新型网络技术将地理分布的算力中心节点连接起来，动态实时感知算力资源调度，进而统

筹分配和调度计算任务，传输数据，构成全局范围内感知、分配、调度算力的网络，在此基础上汇聚和共享算力、数据、应用资源。算力网络包含算、脑(统一感知、编排、调度、协同算力)、网三部分。新型泛在算力和算力网络至少要具备以下特征：

弹性:算力网络的流量特征与互联网的流量特征不完全相同，对于弹性带宽的需求更加突出。例如，在气象的计算场景中，气象中心每天需要计算 1-2 次，每次计算 2 小时，在这 2 个小时内需要非常大的带宽。因此对于气象中心来说，更适合于带宽可调整、时长可定制的弹性连接服务

敏捷:企业客户或者个人用户接入算力网络来获取计算服务，并不需要关心网络中的算力资源和分布情况，只关心算力是否能够敏捷地获取到。

无损:算力由网络来实现互联，网络中的每个丢包，甚至在云数据中心内部的分布式计算过程中的丢包，都会造成算力计算效率的下降。因此，数据中心内部、数据中心之间的无损传输成为算力网络的一个关键特征。

安全:数据是计算的核心要素，也是宝贵资产。安全是算力网络使能到各行各业的一个关键的特征，包括数据安全存储、数据安全加密、算力租户之间数据的安全隔离、外部攻击和数据泄露防护、终端安全接入等。

感知:算力网络中存在海量的应用(算力的需求方)连接，如何为

不同的应用提供差异化的 SLA 保障，又如何为其中重要的应用提供性能的检测和看护，也是算力网络需要考虑的一个关键问题。感知，就是说网络一方面要能够“感知应用”，另一方面还要能够“感知体验综合起来，形成算力网络“应用体验感知”能力。

可视：在算力网络中，需要建立一张网络数字地图，通过应用、算力、网络三者的映射关系和图层建模，形成算(数字世界)和网(物理世界)高效关系映射。网络数字地图对于网络全景进行了动态绘制和动态刷新，可以实现网络拓扑清晰可视、网络路径透明追踪、故障传播关联溯源，以及在算力网络中基于网络、应用、算力关系映射的应用一键导航。

展望未来，各行各业积极把握泛在算网时代的发展机遇，打造高品质网络与泛在的算力供给，提升应用的服务体验，携手产业合作伙伴，不断满足人民对美好信息生活的需要，共创数字经济的美好未来。

参考文献

1.2021-2022 全球计算力指数评估报告，浪潮信息、国际数据公司(IDC)和清华大学,2022

2.中国算力发展指数白皮书, 中国信通院,2021

3.算力网络白皮书，中国移动研究院，2021

4. 云网融合向算网一体技术演进白皮书，中国联通，华为技术有限公司

5. 2022 年算力网络全景洞察白皮书，艾瑞咨询，2022

6. 国家“东数西算”工程背景下新型算力基础设施发展研究报告，国家信息中心，2022

7. 中国联通算力网络实践案例，中国联通研究院，2021

8. 算力网络视频应用白皮书，中国移动研究院，2022

9. 泛在计算服务白皮书 ，中国信息通信研究院，2020

算网融合产业及标准推进委员会（TC621）

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-6230XXXX

传真：010-62304980

网址：www.ccnis.org.cn