中国移动 研究院
China Mobile CMRI

面向智算的算力原生白皮书

（2022 年）

中国移动通信有限公司研究院

前　言

当前，数字经济已成为国民经济高质量发展的新动能，随着人工智能在产业数字化进程中从“单点突破”迈向“泛在智能”，一个以数字化、网络化、智能化为特征的智慧社会正加速到来。智能算力作为人工智能的基石，是算力网络构建多要素融合新型信息基础设施的关键领域，已成为数字经济高质量发展的核心引擎，智能算力基础设施建设也迎来了高潮。

智能算力日益呈现泛在互联的特征，智算生态也呈现碎片化发展趋势，不利于应用的快速创新和算力资源的高效运用，亟需一个可融通业界生态竖井，屏蔽底层算力资源复杂性，提高算力资源利用率，使能应用无感部署、灵活迁移的平台，算力原生技术应运而生。

为凝聚产业共识，进一步推动算力原生技术成熟，中国移动发布本白皮书，分析了智能算力发展的趋势及面临的挑战，系统性介绍了算力原生的定义内涵与愿景、总体架构、关键技术和当前业界的探索实践，并呼吁业界紧密合作、加快构建算力原生统一的标准体系和繁荣的开源生态。

目 录

1. 智算时代：算力原生产生的背景

数字经济时代，随着智慧城市、智慧交通、智慧家庭等智能场景的逐步落地，人工智能正深刻地改变我们的生产、生活方式，一个以数字化、网络化、智能化为特征的智慧社会正加速到来。同时随着5G、边缘计算等支撑技术的持续发展，数智业务转型过程中所产生的数据量正在以更加难以计量的速度爆发，据 IDC 公布的《数据时代2025》显示，从 2016 年到 2025 年全球总数据量将会增长 10 倍，达到 163ZB,其中非结构化数据占 70%以上，计算模式将变得更加复杂，对智能算力的需求也在不断提高，智能计算将成为主流的计算形态。智能计算正以多样化形态广泛融合到生产、生活的各个方面,为千行百业数字化转型提供新动能。

1.1 智能算力呈现泛在互联的特征

依托异构计算、云计算、边缘计算、物联网等技术的持续迭代，算力形态不断演进，呈现泛在化发展特征。在云计算发展的初期和中期阶段，数据中心是算力的主要载体,近年来我国数据中心规模不断增长,工信部发布数据显示，截至 2022 年 6 月,我国在用数据中心机架总规模超过 590 万架,与机架规模增长相适应，数据中心算力规模也在同步增长,我国算力总规模超过 150EFLOPS,位居全球第二。中心化的计算架构提供了集中、大规模的计算、网络和存储等资源，解决了人工智能发展初期面临的业务迅速增长、流量快速扩张，需要大规

模算力的问题。

近些年来，随着云游戏、车联网等新型业务的不断涌现，集中化的云计算服务模式已经无法满足新型业务对数据处理低时延的要求，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，将加速数据处理从云端向边缘的扩散，形成云边端高效协同的发展态势。当前智能算力主要分为大模型训练为主的中心智能算力和以推理为主的边缘智能算力，中心智能算力以智算中心为代表呈现超集中的特征，边缘智能算力以智能边缘云为代表呈现超分布的趋势。边缘算力可以完成覆盖区域内局部业务数据的实时推理和智能决策，减轻云端压力；中心智能算力通过大数据分析，处理优化输出的模型，下发到边缘侧，提升边缘推理准确度。

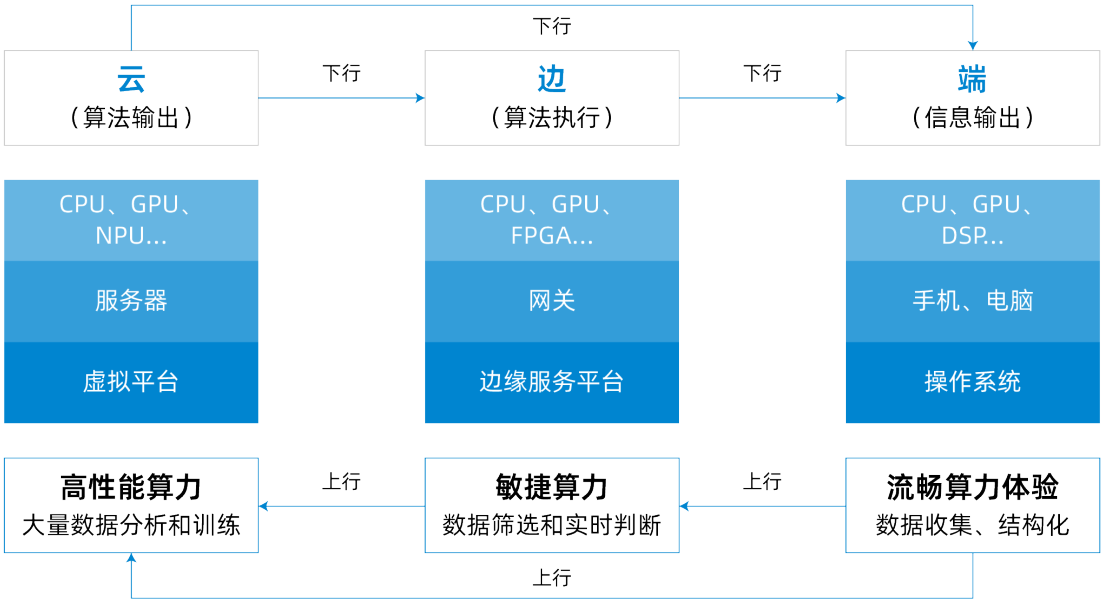

图 1 云边端智能算力高效协同

随着机器学习、自然语义处理等 AI 技术的进一步成熟，AI 训练规模持续扩大、模型复杂度不断提升，模型参数量迈向亿万级，处

理数据量突破千 G 量级，单一智能算力节点已无法满足超大模型训练要求。为使能大模型的分布式训练，加速训练效率，通过以网强算，将各地的智能算力联接成为人工智能算力网络已成为智算产业发展新趋势，产学研各界积极布局，纷纷加入到人工智能算力网络的建设中。2022 年 6 月，鹏城实验室发布“中国算力网（C²NET）”计划，以鹏城实验室的“鹏城云脑”E 级智算平台作为核心节点，广泛联合国家级智算中心、超算中心、大型数据中心以及全国一体化算力网络枢纽节点，大力推进智能算力网络发展。中科院计算所也提出了“信息高铁”科技创新行动计划，将广域环境中的云边端算力资源高速连接，以南京为信息高铁总站，已接入北京高性能站、盐城高通量站、郑州大数据站，并规划接入哈尔滨、太原、西安、合肥、成都五个超算站，提供高通量、高品质和高安全的智能信息服务。

1.2 智能算力生态呈现碎片化发展趋势

人工智能、大数据、AR/VR 等新兴应用的涌现推动了异构计算的迅猛发展，在 intel、NVIDIA、AMD 等传统行业巨头推出 GPU、FPGA 等计算芯片的同时，近年来国内也涌现出燧原、瀚博、沐曦、壁仞、摩尔线程、天数智芯等一批领域化芯片厂商，提供 MLU、NPU、TPU 等领域芯片解决方案，产业一派火热繁荣景象。各厂商为使能各自芯片在更广泛领域运用，吸引更多的应用开发者，在提升芯片本身性能的同时不断围绕自身芯片架构构筑各自的软件生态，包括编译工具链、操作系统适配组件以及相关核心软件库等。然而，各厂商的芯片架构

千差万别，因此其驱动、软件开发接口、软件运行时等各不相同且难以兼容，同时，各芯片厂商会尽可能在其芯片工具链中融入TensorFlow、PyTorch 等各类 AI 框架，并结合各自芯片特点进行针对性优化定制，进而形成多个厂商的分支版本，不同分支版本之间的代码难以相互移植，更导致了各厂商软件生态的碎片化和竖井化。

1.3 泛在多样的智能算力发展面临的挑战

如何将遍布在云边端泛在部署的异构多样算力资源以及繁多碎片化软件生态间进行有效协同，驱使业务应用能平滑的在各级算力资源上进行流转运行，充分利用巨量算力资源，是使能智算业务转型创新的关键点。

一是对开发者来说，要实现跨架构的应用优化部署，开发成本高。应用开发人员在使用异构算力进行 AI 算法实现过程中可以很明显地感受到，不同类型的 AI 处理器的应用程序接口、编程库和操作系统服务不是统一的，例如 CUDA、Vitis、MindSpore 等，以 ASIC 为主的专用芯片的领域编程范式和工具链更是种类繁多，而且目前没有一种编程模型能够适用于所有异构系统，需要在 OpenCL、OpenACC、OpenMP 等多种模型范式间切换。这些竖井式的开发生态增加了代码开发成本，一名开发人员很难精通多类硬件特性及开发环境，为了开发出能够适配多种异构算力的应用程序，企业需建立多支开发团队、维护多个程序版本，带来巨大的开发成本，这已经成为协同运用多样算力的一个主要瓶颈问题。

二是对算力服务商来说，无法实现异构算力的合理规划和应用的动态迁移，资源利用率低。首先，算力服务商通常会基于多个厂商的多类 AI 服务器进行建设，不同厂商不同架构的 AI 服务器资源池互相独立，生态隔离，形同一个个孤岛。算力服务商在资源规划时需要结合多种硬件上预估的应用规模进行采购，很难实现精准配比，同时由于上层应用与底层硬件紧绑定关系，后期也无法实现迁移调整，因此可能会出现某些厂商的硬件资源不够而另外一些厂商的硬件资源闲置的状况；其次，当前 GPU、AI 芯片虚拟化能力存在局限性，物理资源只能以独占式的分配方法提供给用户实例使用，无法实现动态调整和灵活调度，导致底层资源无法被充分利用。以 GPU 为例，当前只支持几种固定份数切割的 vGPU 虚拟化能力，且用户实例对 vGPU 的挂载是独占式的，如果用户实例需要增加或减少 vGPU 数量，还需要对实例进行重启，这在流量高峰期可能会导致业务中断。

三是对新兴芯片制造商来说，面对当前逐步形成的一超多雄的产业格局，良性发展的生态构建难。AI 芯片从制造到大规模应用，还隔着一个巨大的产业和生态鸿沟，一是要与基础软件生态适配，构建可发挥自身特性的驱动程序、编程工具链以及 API 接口等初始软件能力更重要的是要实现与算法框架和 AI 应用的优化与适配，“AI 框架+AI 芯片”的组合在一定程度上决定了产品的技术路线和市场受众。当前，95%以上的智能应用均构筑在 Tensorflow、Pytorch 等国际主流 AI 框架之上，而这些框架从底层代码、接入机制，到中间算子的迭代研发，均由头部巨型 AI 芯片厂商参与主导，不断针对自身处理

器特性进行优化，且以官方版本发布，在用户受众中已形成事实标准。新兴 AI 处理器厂商面临两难境地，一方面是难以有机会参与到 AI 框架的建设工作，无法获得官方发布支持；另一方面是其所建立的 AI 框架分支版本，由于受众群体小，开发者更是少之又少，缺乏应用移植迁移技术，难以承接到支持其技术迭代发展的足够业务量，常常在生存边缘徘徊挣扎，难以发展壮大，陷入“差而不用、不用更差”的怪圈。

2. 算力原生定义内涵与愿景

2.1 算力原生定义内涵

算力狭义上是指多样计算器件所构成的设备、集群、平台等算力资源对数据的处理能力。原生则意味着事物最原始、最本真的状态。

为解决智能算力发展面临的一系列挑战，算力原生**回归计算本真**，是以**统一的算力资源抽象模型**和**标准的编程范式接口**为基础，以**跨架构编译优化**技术和**原生运行时**技术为依托，使能多样泛在算力环境下屏蔽复杂软硬件差异的技术，实现同一应用一套代码、动态重构一体部署、灵活迁移高效执行。

2.2 算力原生愿景

算力原生通过构建标准统一的算力抽象模型及编程范式接口，打造开放灵活的开发及适配平台，实现各类异构硬件资源与计算任务有效对接、异构算力与业务应用按需适配、灵活迁移，充分释放各类异构算力协同处理效力、加速智算应用业务创新，**实现异构算力资源一体池化、应用跨架构无感迁移、产业生态融通发展的目标愿景**。

一套代码，简化开发。应用开发者面向底层多样异构的算力芯片，无需针对性开发多套代码并进行优化、集成、验证工作，在不改变编程语言的前提下，只须一次开发即可在多种芯片架构上运行。

异构资源，一体池化。形成多厂商、多架构的异构智能算力混合资源池，实现从传统的以硬件资源为单位、静态分配使用算力的方式，

变为以计算能力为单位对算力资源进行动态、灵活地配给，应用无需关注智能算力的位置、数量和类型。

屏蔽差异，跨芯迁移。屏蔽各厂商多架构智算芯片的软硬件差异，高效生成可跨架构流转、任务式互映射的原生程序，应用可实现跨架构的无感迁移与协同部署。

融通产业，繁荣生态。破解当前智算产业生态碎片化、割裂化问题，打通智算产业生态融通渠道，形成百花齐放的产业繁荣新生态和发展新格局。

3. 算力原生平台架构与关键技术

3.1 算力原生平台架构

算力原生平台的任务目标是深度适配泛在、多样的异构算力资源，构建标准开放的算力原生技术栈，使能上层应用一次开发、跨架构无感适配执行、任意流转迁移。算力原生技术架构包括：算力池化层、算力抽象层。

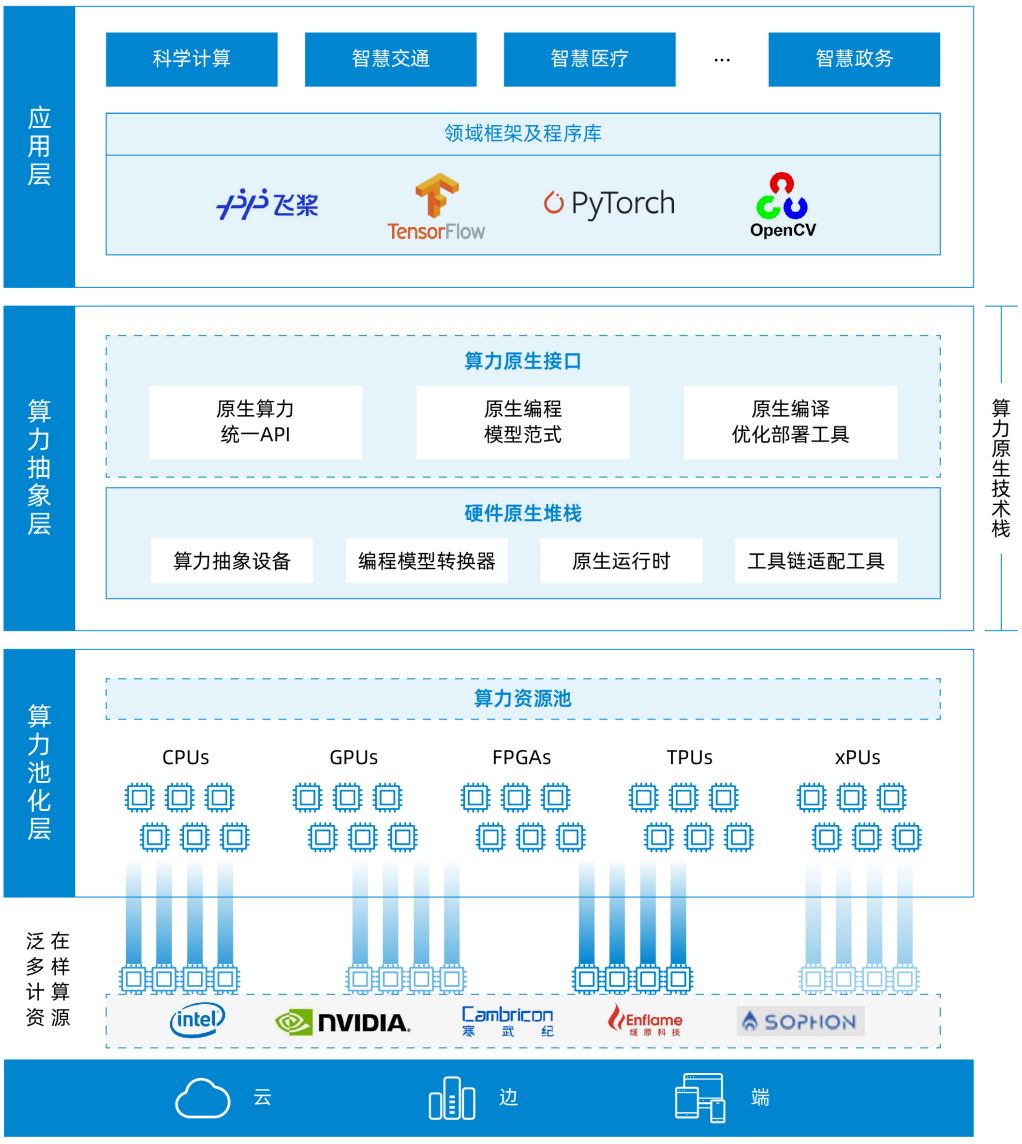

图 2 算力原生平台架构

（1）算力池化层

算力池化层通过构建底层异构硬件的统一抽象模型，并对应用调用底层算力资源的请求进行重定向和再调度，从而实现各类硬件资源的一体池化，从传统的以硬件资源为单位、静态分配和调度的方式，变为以计算能力为单位对算力资源进行动态、灵活地配给。同时为应对智算业务的潮汐效应，算力池化层可根据业务需求及算力负载情况提供算力资源弹性扩缩容的能力。

（2）算力抽象层

算力抽象层由硬件原生堆栈和算力原生接口组成，**其中硬件原生堆栈**主要包括编程模型转换器和原生运行时，编程模型转换器可将基于特定芯片编程的应用程序转译为与底层硬件架构无关的算力原生中间元语；原生运行时可实现对底层算力资源的感知和控制，完成原生程序的加载、解析，保障计算任务与本地计算资源的即时互映射，按需执行。

算力原生接口基于原生算力抽象接口及多模混合并行编程模型，构建原生算力统一API、原生编程模型范式及原生编译优化部署工具，形成可嵌入式融入用户业务的开发环境，辅助用户生成可跨架构流转、无感迁移与任务式映射执行的算力原生程序。

3.2 算力原生演进路径

算力原生将从**实现异构算力资源池化、实现应用的跨架构迁移、全局泛在融通**三个阶段分步推动技术成熟，破解竖井式异构生态发展

难题，实现异构算力资源一体池化，使能应用跨架构无感迁移。

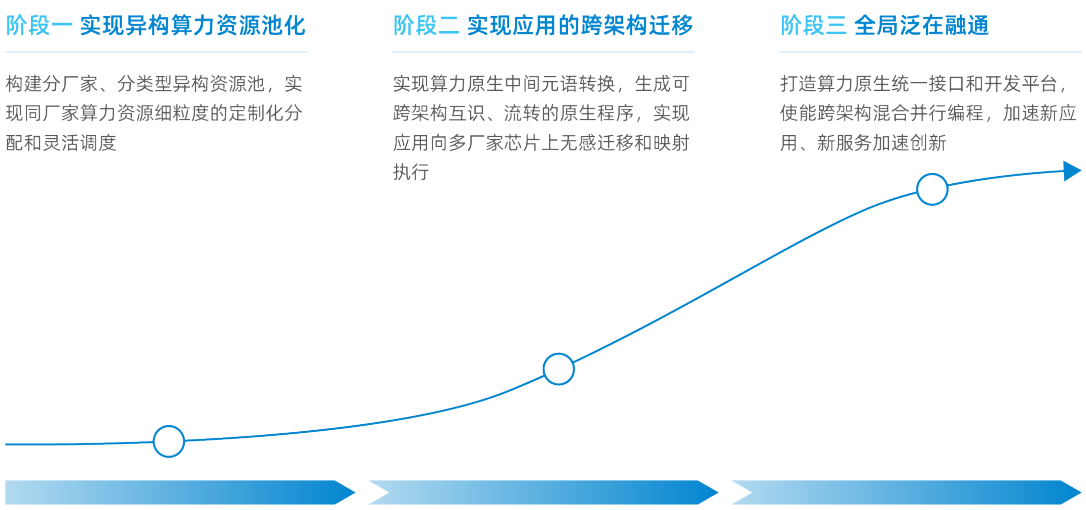

图 3 算力原生演进路径

阶段一：核心目标是实现异构算力资源池化。通过构建分厂家分类型的异构资源池，引入对应用调用底层算力资源 API 的重定向技术，实现同厂家算力资源细粒度的定制化分配和灵活调度。在本阶段可实现应用对底层算力资源位置和数量的无感使用，支持业务可根据动态负载情况进行算力资源的弹性使用，充分应对潮汐效应和业务量突发等场景。

阶段二：核心目标是实现应用的跨架构迁移。在实现分厂商、分类型资源池化的基础上，构建虚拟资源模型、屏蔽底层硬件差异，实现跨厂商、跨架构的硬件资源池化供给。同时通过引入编程模型转换器，可实现基于特定芯片厂家开发的应用，在用户无需重构代码的基础上，转换成算力原生中间元语，结合算力原生运行时和各厂商工具链，生成可跨架构互识、流转的算力原生程序，最终实现在多厂家芯片上无感迁移和映射执行。

阶段三：核心目标是实现全局泛在融通。为支撑算力网络中远期

实现算力解耦、泛在调度的任务式服务目标，算力原生基于一二阶段实现池化和跨架构无感迁移的基础上，通过打造算力原生统一接口和开发平台，形成统一编程模型及开发环境，使能应用开发者实现跨架构的混合并行编程，加速新应用、新服务创新。

3.3 重点攻关方向与关键技术

3.3.1 算力抽象及异构算力统一编程模型技术

算力抽象旨在屏蔽应用对底层多样硬件架构的差异感知，通过抽象化的设计建立一套支持多种异构算力系统的编程模型，是异构算力资源对上层应用及开发者之间的桥梁，一方面使开发者在编程时既可以充分利用丰富的异构资源、又不必考虑复杂的系统细节，同时，使基于抽象模型所开发出的应用程序在运行时可以根据系统架构的变化，即时转换映射，实现快速迁移移植。算力抽象既能够包容多样算力异构，提供一个统一编程模型，又能够保持异构算力自身的并行效率，将是异构混合计算编程的重要发展趋势。

然而，与传统同构系统相比,异构算力系统的统一编程模型建立面对如下挑战：

（1）针对不同异构算力内核的指令集、并行模式、任务映射和并行性能存在极大差异、属性不一的现状，寻求能够广泛涵盖多样异构算力特性的统一编程模型及范式极其困难；

（2）算力内核在内部结构、互联架构两方面都变的更加复杂化和多样化，随着算力核心数目的高动态扩缩容，带来了复杂的计算任

务并行机制设计挑战；

（3）算力核心数目的弹性扩缩容导致对计算任务所需缓存、I/O带宽的要求也不断变化，传统共享内存机制下的数据协同无法满足性能需求，而设计多层次、多模态的复杂数据缓存结构及数据同步机制将极赋挑战。

为应对上述挑战，应着重在如下方面开展技术攻关工作：

（1）针对异构算力系统中各加速单元间并行计算能力、缓存资源不同的问题,探索在传统并行编程模型的基础上增加“异构特征描述”范式，演进实现异构任务划分机制，用于描述任务在不同算力单元间的分配；

（2）针对异构算力系统中加速设备内数据分布可重构、算力单元间缓存数据交互渠道多样的问题,探索在传统并行编程模型的基础上增加多层数据分布描述范式，拓展现有异构计算中共享数据模型；

（3)针对异构算力系统中大范围、多类型数据同步操作的问题，探索在传统并行编程模型的基础上增加“算力单元间同步”机制,同时,依据加速设备硬件特征,提供算力系统内的局部和全局即时巨量同步。

3.3.2 算力原生接口及异构算力编译优化技术

开发面向异构算力系统的算力原生接口能够极大的减少开发者的编程负担。并行编译器能够发现串行程序中的可并行性、自动的并行循环和向量化数据优化操作，将异构化、并行化的工作留给编译器

可以显著降低编写并行程序的时间及人力成本，同时也将极大提高所生成算力原生程序的可移植性。

同构系统的编译过程相对简单，然而对于异构算力系统，由于协处理器设计简单，更多的硬件细节要交给软件处理，因此对编译技术也提出了更高的要求。

（1）**攻关面向异构算力的原生代码自动生成技术。**基于统一编程模型及范式，探索源到源编译器，利用数据对齐、数据分布等技术将程序中的数据自动划分到不同的处理器核中，并根据不同数据分布机制为应用程序插入通讯原语，生成可多模重构的算力原生程序；

（2）**攻关面向异构算力的数据自动管理技术。**针对 GPU、MLU 等加速核局部存储器容量无法满足智算中大数据规模的训练运用的问题，应攻关实现数据自管控系统，该系统通过分级数据分布、通讯生成和循环分块等方法对程序中的数据和计算进行分解，使得分解后的数据能够满足局部存储容量的约束。

3.3.3 硬件原生堆栈及运行时支持机制

统一编程模型通过算力原生接口呈现给开发者，将开发者编写的源程序编译为可执行文件,最终通过运行时系统完成任务的执行。运行时支持机制的主要任务是保障任务映射，即任务具体在哪个算力单元上执行、以何种顺序执行。同时还负责对任务划分、数据分布与通信、同步等机制进行全系统级别的优化。

硬件原生堆栈及运行时系统的任务映射机制有两类:一是直接映

射，即独立完成并行任务向异构平台映射的工作；二是间接映射，即需要借助其他异构编译/运行时系统协助完成部分任务映射工作。直接映射机制通常在运行时系统中实现，而间接映射机制则采取编译时源到源变换与运行时分析相结合的方式实现。

为解决异构算力系统中设备间并行计算能力存在差异的问题，重点研究异构计算任务再映射机制，研究如何将任务自动地分配到GPU、MLU、多核 CPU、定制类 SIMD 加速 ASIC 系统等不同算力单元上执行。

为了解决异构算力系统中加速设备内数据分布可重构、设备间数据通信渠道多样、同步范围复杂的问题，需要开展运行时在细粒度计算任务的分配、多种数据存储与通信、共享与同步等方面的研究工作。

4. 算力原生产业实践

4.1 业界厂商实践

当前，算力原生已成为产业界重点关注和大力攻关的热点技术领域，阿里云、英特尔、驱动科技等厂商对算力原生的部分关键技术进行了攻关探索，已取得了一定进展。

阿里云在 AI 场景下针对统一异构硬件接口和编译平台方面进行了探索。针对业界缺少统一的异构硬件接口标准以及 AI 芯片厂商需对不同框架进行适配的问题，阿里云提出了基于 HALO/ODLA 的 AI 部署解决方案，异构计算编译框架 HALO （Heterogeneity Aware Lowering Optimization）可以基于面向深度学习的异构硬件统一接口规范 ODLA（Open Deep Learning API），将模型编译为与 AI 框架、设备无关的代码，应用经 HALO 与厂商提供的 ODLA 运行库链接后，就可运行在相应的平台上。这种方式，无需厂商提供底层的指令集或编译后端，只需提供符合接口的运行库。同时厂商基本都会提供算子级别的运行库，与 ODLA 对接容易，对新硬件的支持比较友好。目前除了 CPU，GPU 外，已经支持了 Graphcore IPU, 寒武纪思元 370，云天励飞 DeepEye 1000, 高通 AIC 100 等多款硬件。

英特尔公司在跨平台统一开发和打造开源生态方面进行了探索。为满足新型应用在复杂场景、多任务并行和多计算架构组合的有效协同，英特尔推出了一个跨架构编译规范 OneAPI。它是一种跨行业、开放、基于标准的统一编程模型，提供了通用、开放的编程体验，让

开发者可以自由选择架构，无需在性能上作出妥协，大大降低了使用不同的代码库、编程语言、编程工具和工作流程带来的复杂性。同时 OneAPI 规范还基于现有标准创建了 OneAPI 组件规范的开源实现，以使 OneAPI 能够快速用于新的硬件架构或软件语言。

趋动科技在计算资源池化方面进行了探索。为解决独占式 AI 算力资源，降低 AI 算力使用成本，趋动科技推出了猎户座（OrionX） AI 算力资源池化解决方案。该方案颠覆了原有的 AI 应用直接调用物理 GPU 算力的方法，通过在 AI 应用与物理 GPU 之间增加软件层实现两者解耦，并通过应用调用 AI 算力的 API 重定向技术，实现 AI 算力资源的池化和细粒度分配。

4.2 中国移动“芯合”算力原生原型系统

中国移动联合阿里云、趋动科技等产业合作伙伴，构建了“芯合”算力原生原型系统，可以提供算力资源池化及细粒度配合、原生编译及跨架构部署迁移两个方面的能力。

（1）算力资源池化及细粒度配给

算力原生原型系统具备将云、边异构算力资源分厂商、分类型池化的能力，当前已实现英伟达 GPU、寒武纪 MLU 的池化和细粒度调度，可针对应用的业务负载情况进行异构算力资源的弹性扩缩容，应用无需感知底层异构算力资源的位置和数量，即可提升算力资源池整体的利用率。

（2）原生编译及跨架构部署迁移

在对 GPU、MLU 进行池化的基础上，算力原生原型系统还具备对上述异构算力资源进行抽象和屏蔽的能力，当前已实现基于深度学习的图像识别、视频流分析等智算应用在英伟达 GPU 和寒武纪 MLU 上跨架构迁移部署，使能上层应用并对底层异构算力资源无感使用。

4.3 算力原生开源建设

为打造开放共享的开源生态环境，推动算力网络技术的稳步健康发展，中国移动于 2022 年 7 月在开放基础设施基金会（OIF，OpenInfra Foundation）主导发起和成立了全球首个算力网络开源社区（CFN，Computing Force Network Working Group）。算力原生作为算力网络开源工作的重要攻关领域，成为了工作组首批成立的子工作组之一，并得到了包括中国移动、阿里云、浪潮、九州云在内的众多产业界合作伙伴的积极响应和参与。后续，算力原生子工作组将会在统一算力抽象模型、算力原生编译平台、标准原生系统接口和算力原生运行时四个部分进行重点突破，期望能够输出全球首个跨架构算力原生开源平台代码实现，同时欢迎产业界合作伙伴积极参与，共筑算力原生产业生态，推动算力原生技术的成熟落地。

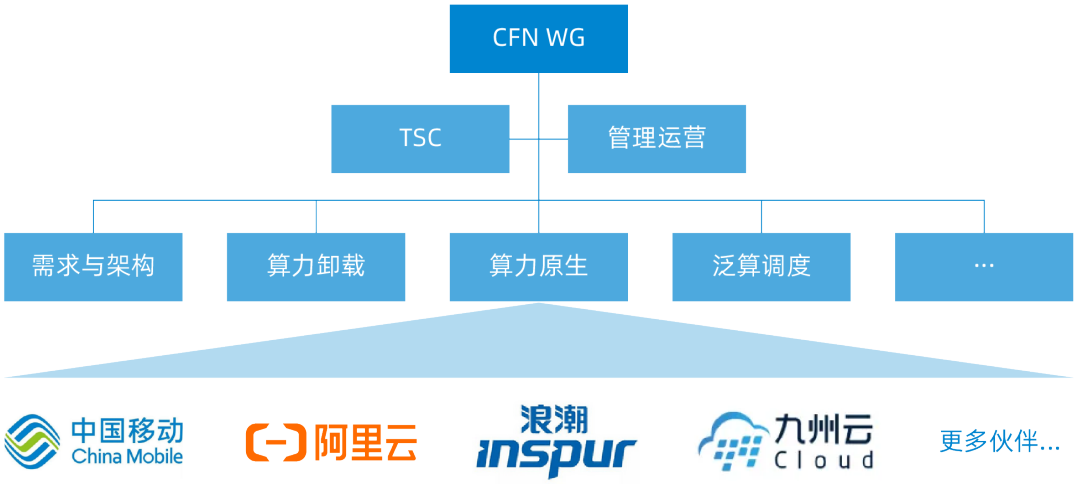

图 4 算力网络 CFN 开源社区算力原生子工作组

5. 展望与倡议

当前，中国移动已经开展了算力原生关键技术的研究，为推动算力原生技术成熟，**实现异构算力资源一体池化、应用跨架构无感迁移、产业生态融通发展的目标愿景**，中国移动呼吁产学研各界合作伙伴精诚合作，凝聚共识，共同推进算力原生技术成熟，繁荣产业生态，提出以下几点倡议：

联合展开算力原生关键技术攻关。联合攻关算力抽象技术、统一编程模型技术、算力编译优化技术、算力硬件堆栈及运行时支持技术，基于中国移动算力网络试验示范网项目，开展算力原生原型的试验试点验证工作。

联合推动算力原生标准化建设。联合制定标准化的算力抽象统一接口，联合推进算力原生在 ITU、CCSA 等组织的标准化工作，拉通软硬件开发生态，为推动构建面向全球的标准化统一算力原生开发平台打下坚实基础。

联合打造算力原生开源社区。联合众多产业界合作伙伴，在算力网络开源社区成立算力原生开源社区，输出全球首个跨架构算力原生开源平台代码实现，欢迎产业界积极参与，共筑算力原生开源开放的产业生态。

参考文献

[1]算力网络白皮书[R]，中国移动，2021

[2]算力网络技术白皮书[R]，中国移动，2022

[3]中国算力发展指数白皮书[R]，中国信息通信研究院，2021

[4]AI 框架发展白皮书[R]，中国信通院，2022

[5]新型数据中心发展三年行动计划（2021-2023 年），中国工业和信息化部，2021

缩略语列表

缩略语	英文全程	中文释义
HALO	Heterogeneity-Aware-Lowering-and-Optimization	异构感知编译优化
ODLA	Open Deep Learning API	面向深度学习通用加速硬件的开放接口
GPU	Graphics Processing Unit	图形处理器
NPU	Neural Processing Unit	神经网络处理器
IPU	Intelligence Processing Unit	GraphCore AI 加速芯片代号
xPU	“x” Processing Unit	各种AI加速硬件的统称，如:GPU，IPU，NPU等
HIC	Heterogeneity Intelligence Card	异构加速卸载卡
ASIC	Application Specific Integrated Circuit	特殊应用集成电路
FPGA	Field Programmable Gate Array	现场可编程逻辑门阵列
IDE	Integrated Development Environment	集成开发环境
SIMD	Single Instruction Multiple Data	单指令多数据流
CUDA	Compute Unified Device Architecture	显卡厂商NVIDIA推出的运算平台
API	Application Program Interface	应用程序接口
OpenCL	Open Computing Language	开放运算语言
OIF	OpenInfra Foundation	开放基础设施基金会
CFN	OpenInfra Foundation	算力网络开源社区
GPGPU	General-purpose computing on graphics processing units	通用图形处理器
IDC	Internet Data Center	互联网数据中心
ITU	International Telecommunication Union	国际电信联盟
CCSA	China Communications Standards Association	中国通信标准化协会
AR/VR	Augmented Reality/Virtual Reality	增强现实/虚拟现实
EFLOPS	Exa Floating-point Operations per Second	每秒一百京次浮点运算


中国移动
China Mobile
研究院
C M R I