算力度量与任务调度：物联网端侧设备策略研究

祝淑琼，徐青青，李小涛，陈维
（中国移动通信有限公司研究院，北京 100053）

摘　要：随着移动通信、人工智能等技术的发展，智能设备和数据呈现爆炸式增长的态势，物联网（Internet of things，IoT）场景对算力、时延和能耗提出了更高的要求。算力网络通过对计算节点进行互联，基于统一的算力度量标准和任务调度策略，实现算力资源的共享和高效利用，为提升物联网系统的计算性能提供了新思路。但是由于物联网设备种类繁多、网络连接方式各异且对功耗敏感，当前以计算能力为主的算力度量方法无法满足物联网设备协作的需求。此外，目前算力网络的计算任务调度方法普遍依赖中心化的网络路由节点或管理平台，不能适应物联网设备分布离散和资源受限的特点。针对上述问题，提出了一种面向物联网端侧设备的新型算力度量架构，为异构物联网端侧算力资源提供计算、存储、通信、功耗和电源的统一度量。在此基础上，提出了一种分布式的任务调度策略，实现离散异构算力资源与业务场景需求的智能匹配，支持物联网端侧设备的资源管理和任务调度。选取智慧家庭场景对提出的算力度量架构进行评估，结果表明，该架构可以有效地实现端侧设备算力资源的共享和调度，提升物联网计算效率，减少能源消耗。

关键词：算力网络；算力度量；物联网；任务调度
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024084

Computational measurement and task scheduling: a study on IoT edge device strategies

ZHU Shuqiong, XU Qingqing, LI Xiaotao, CHEN Wei
China Mobile Research Institute, Beijing 100053, China

Abstract: The rapid advancement of mobile communications and artificial intelligence has catalyzed an exponential increase in intelligent devices and data generation. This surge necessitates enhance the computational resource capabilities, particularly in the Internet of things (IoT) environments, where there are pressing demand for improved resource management in terms of computation, latency, and energy efficiency. The concept of a computility network, which leverages interconnected computing nodes for resource sharing and optimization based on a unified measurement standard and task scheduling strategy, offers a promising solution for augmenting IoT systems′ computational performance. However, the current models for computing resource measurement, predominantly focused on computational capacity, fall short in addressing the diverse and collaborative needs of various IoT devices. These devices often

收稿日期：2023−12−04；修回日期：2024−03−28
通信作者：祝淑琼，zhushuqiong@chinamobile.com

differ in network connectivity modes and exhibit sensitivity to power consumption. Moreover, prevalent task scheduling methods in computility network predominantly rely on centralized network routing nodes or management platforms. Such approaches are not well-suited for the unique characteristics of IoT devices, which are typically dispersed and constrained in resources. To address these challenges, a novel architecture for computing resource measurement tailored to IoT devices was introduced. A comprehensive and unified framework for measuring diverse aspects of computing resources in heterogeneous IoT environments was provided, including computation, storage, communication, power consumption, and power supply metrics. Building on this foundation, a distributed task scheduling strategy that intelligently aligned the disparate computing resources with specific business scenario requirements was proposed, thereby facilitating efficient resource management and task scheduling for IoT devices. To validate the effectiveness of the proposed architecture, it was applied to a smart home scenario. The empirical results demonstrate that the proposed architecture significantly enhances the sharing and scheduling of computing resources among IoT devices. It elevates the overall efficiency of IoT computing while concurrently reducing energy consumption, thereby offering a robust solution to the evolving demands of IoT systems.

Key words: computility network, computing resource measurement, IoT, task scheduling

0　引言

算力指计算机系统或设备所具备的性能和能力，包括处理器速度、存储器容量、通信带宽等方面。算力已成为人工智能、大数据、物联网等领域发展的核心驱动力。根据 IoT Analytics 发布的研究报告，2022 年全球物联网连接数已达到 143 亿，预计到 2027 年，全球物联网设备的连接数量将超过 290 亿。高速增长的物联网设备不断拓展应用场景，深刻影响人们的生活方式。与此同时，随着移动通信、人工智能等技术的迅速发展，物联网场景对算力、时延和能耗提出了更高的要求[1]，单一的物联网设备逐渐无法满足业务需求。算力网络[2-3]作为实现云网边端业深度融合的新型信息基础设施，通过统一的算力度量标准实现对算力资源的评估管理，并通过任务调度策略实现算网资源的动态分配，最终提升算力资源的利用率。算力网络为提升物联网系统的计算性能，丰富物联网的业务场景提供了创新思路。

目前算力网络中的算力度量架构以计算和存储能力为主，包括对中央处理器（central processing unit，CPU）、图形处理器（graphic processing unit，GPU）、神经处理单元（neural processing unit，NPU）等芯片的计算性能和内部存储资源的评估。由于物联网端侧设备普遍是能耗敏感型设备，并且部署环境千差万别，通信方式多种多样，现有算力度量方法缺乏对设备电源能力和通信能力的量化评估，无法满足物联网算力度量的需求。此外，为了实现算力资源和计算任务的匹配，需要对计算任务进行按需调度，但当前算力网络中，计算任务的调度过程普遍依赖中心化的网络路由节点或管理平台，对于执行任务调度的设备的算力有很高要求，与物联网设备分布离散、资源受限的特性相违背。

针对上述问题，本文提出了一种面向物联网端侧设备的新型算力度量架构。该架构包含 3 个核心层面：算力资源统一度量层，任务需求准确度量层和高效计算任务调度层。首先，算力资源统一度量层针对异构物联网端侧设备的算力资源（包括计算、存储、通信、功耗和电源 5 个维度）进行全面和统一的度量，以满足不同物联网业务场景的多样化需求。这一层的设计旨在为各类物联网设备提供一个综合的评估框架，以确保资源利用的最大化和业务执行的效率。其次，任务需求度量层对计算任务所需的算力资源和服务质量进行细致评估。这一层支持算力资源与计算任务

之间的高效匹配，确保每个任务都能在最佳的设备上执行，从而优化整体的服务质量和执行效率。最后，基于上述两层的度量结果，提出了一种分布式任务调度策略。该策略利用服务能力评估模型，实现离散且异构的算力资源与业务场景需求之间的智能匹配。这种策略支持物联网端侧设备在资源管理和任务调度方面的高效运作，有助于提升整体的计算效率和实时性。为了验证所提架构的有效性，本文以智慧家庭应用为案例进行了实证评估。结果表明，该架构能够有效地实现端侧设备算力资源的共享和调度，提升物联网业务的计算效率和实时响应能力。综上所述，本文的主要贡献如下。

（1）提出了一种全面的算力资源统一度量方法，为物联网端侧设备提供全方位的评估维度和具体的度量指标，满足了复杂物联网环境下的度量需求。

（2）提出了一种细致的任务需求度量方法，能够准确评估执行计算任务所需的算力资源和达到的服务质量。

（3）提出了一种创新的分布式任务调度策略，通过智能匹配，实现了物联网设备的算力资源和任务需求间的有效对接。

1 相关工作

算力网络旨在对算力节点进行组织和管理，基于统一的算力度量标准和灵活的任务调度策略，提供随取随用的算力服务，并提高网络和计算资源的利用率[4-5]。国际电信联盟电信标准化部门（ITU-T）第13研究组（SG13）在2021年7月发布了首个算力网络领域国际标准Y.2501[6]，标准规定了算力网络的功能架构。中国通信标准化协会（CCSA）网络与业务能力技术工作委员会（TC3）发布了《算力网络总体技术要求》[7]和《面向算网融合的算力度量与算力建模研究》[8]等相关标准。当前，业界工作主要集中在算力网络整体功能和技术框架的研究，对算力度量和调度策略的研究仍处于初级阶段，尚未形成统一的规范和标准。

1.1 算力度量方法

目前，对于算力度量的研究主要集中在单芯片算力以及数据中心或计算节点的综合算力的度量。文献[9]将芯片算力分为逻辑运算能力（CPU）、并行计算能力（GPU）和神经网络计算能力（NPU和张量处理单元（tensor processing unit，TPU）），同时提供了一种考虑各类芯片数量的度量映射函数，将异构的算力映射到统一量纲。文献[10]考虑了设备的计算资源动态使用特性，不仅度量了设备本身具有的计算资源（CPU和GPU）最高每秒能完成的计算指令次数（million instructions per second，MIPS），还将计算资源的空闲率纳入算力度量维度。文献[11]提出了一种数据中心能效度量方法，测算了我国当前的数据中心算力和能效水平。但其只考虑了浮点计算能力，不能满足一些低精度计算的物联网设备算力评估需求。文献[12]提出了一种基于端边云的算力度量方法和架构，根据采集的算力资源信息从CPU、内存和网络利用率3个维度构建算力度量矩阵，并针对端边云不同业务类型赋予不同权重值，形成业务应用与算力资源的对应。该方法考虑了端边云不同类型业务的关联问题，但是由于算力资源匹配不同类型业务的能力赋予权重仅考虑历史经验来人为设定，存在主观性较强的问题，无法实现算力资源和任务的精准匹配，并且存在任务调度失败的风险。文献[13]提出了“四面三级”算力度量技术体系，从节点的计算、通信、内存和存储能力4个方面来建模算力网络的异构算力。但是该方案没有考虑物联网设备存在功耗敏感、网络连接方式多样和算力资源受限的特征。

1.2 算力网络中的任务调度方法

任务调度研究方法以算力资源和任务需求匹

配为主。文献[14]提出了一种协同调度管理云边端算力资源的算力网络架构，包括计算资源池、算力网关和网络3个部分。根据资源池CPU计算能力、内存大小和硬盘存储空间，计算任务通过网络调度至资源池中完成计算需要的时间的建模方法，从而获取不同资源池和任务完成时间的匹配情况。文献[15]通过对计算任务和算力资源从时间复杂度、空间复杂度、任务计算类型（CPU、GPU、NPU）进行量化，构建马尔可夫决策过程对任务执行时间进行预测优化建模，得到最优资源分配策略。但是这些方法不仅计算复杂，也忽视了能耗的影响。文献[16]提出了一种算力服务需求匹配系统，系统中算力网络解析服务器根据算力资源管理服务器获取实时算力资源信息，将用户的算力需求匹配到算力资源节点。文献[17]提出了一种混合式路由调度方案，通过算网编排层实现了算力状态的感知，利用算力路由设备完成了算力与网络一体的最优路径计算和服务调度。但是以上方法均依赖中心化的大算力路由节点或管理节点，并不适用于设备分布离散、资源受限的物联网应用场景。

综上所述，当前算力度量方法和任务调度方法无法满足设备种类繁多、网络连接方式各异、对功耗敏感且设备资源受限的物联网端侧设备算力资源度量和计算任务调度需求。针对以上问题，本文提出了面向物联网端侧设备的算力度量方法，相比现有方法新增了设备计算精度、通信方式、功耗情况、电源能力等的度量指标，满足了物联网端侧设备算力资源的度量需求。此外，本文面向实际业务进一步将服务质量纳入度量架构，对计算任务的需求进行了全面评估，选取了合适的设备来完成该任务的计算。在任务调度方面，与现有基于中心式节点的调度方法不同，本文提出了一种分布式算力调度方法，由需要卸载任务的物联网端侧设备广播需求，闲置设备响应请求并反馈自身的算力资源信息，根据闲置算力资源的度量信息以及需要调度的计算任务需求度量情况，完成任务和算力资源的匹配判断，从而实现高效的任务调度。与现有方案相比，本文方法计算复杂度更低，并且可以由端侧设备独立完成调度，更符合物联网端侧设备分布离散及资源受限的特点。

2　物联网算力度量架构

物联网终端算力作为算力网络的关键部分，通过海量的终端设备给多样化需求的物联网业务场景提供泛在的算力服务。但由于物联网端侧设备表现出显著的算力异构性、多样的通信方式、对功耗的高度敏感及资源的有限性，现有的算力度量架构无法满足物联网端侧设备在算力度量和任务调度方面的要求。

2.1　总体架构

为了使物联网端侧设备能够通过统一的算力度量方法实现算力资源共享，并且通过任务调度满足不同业务的算力资源需求和服务质量需求，同时针对不同业务需求进行有效的任务调度，以提供更高质量的应用服务，本文提出了一种新型的面向物联网端侧设备的算力度量架构，如图1所示，该架构由算力资源统一度量层、任务需求度量层和计算任务调度层组成。

2.1.1　算力资源统一度量层

此层不仅包括传统算力度量方法中的计算芯片类型、存储能力、传输带宽等指标，而且针对物联网端侧设备的特点，如算力异构性、多样化通信方式、功耗敏感性，增加了计算精度、通信方式、功耗及电源能力等评估指标。这些补充指标使算力度量体系更加全面，能够更好地满足物联网端侧设备的算力资源评估需求。综合考虑这些多维度指标，算力资源统一度量层为物联网设备提供了一个全面、精准的评估框架，确保资源配置的最优化和业务需求的高效满足。

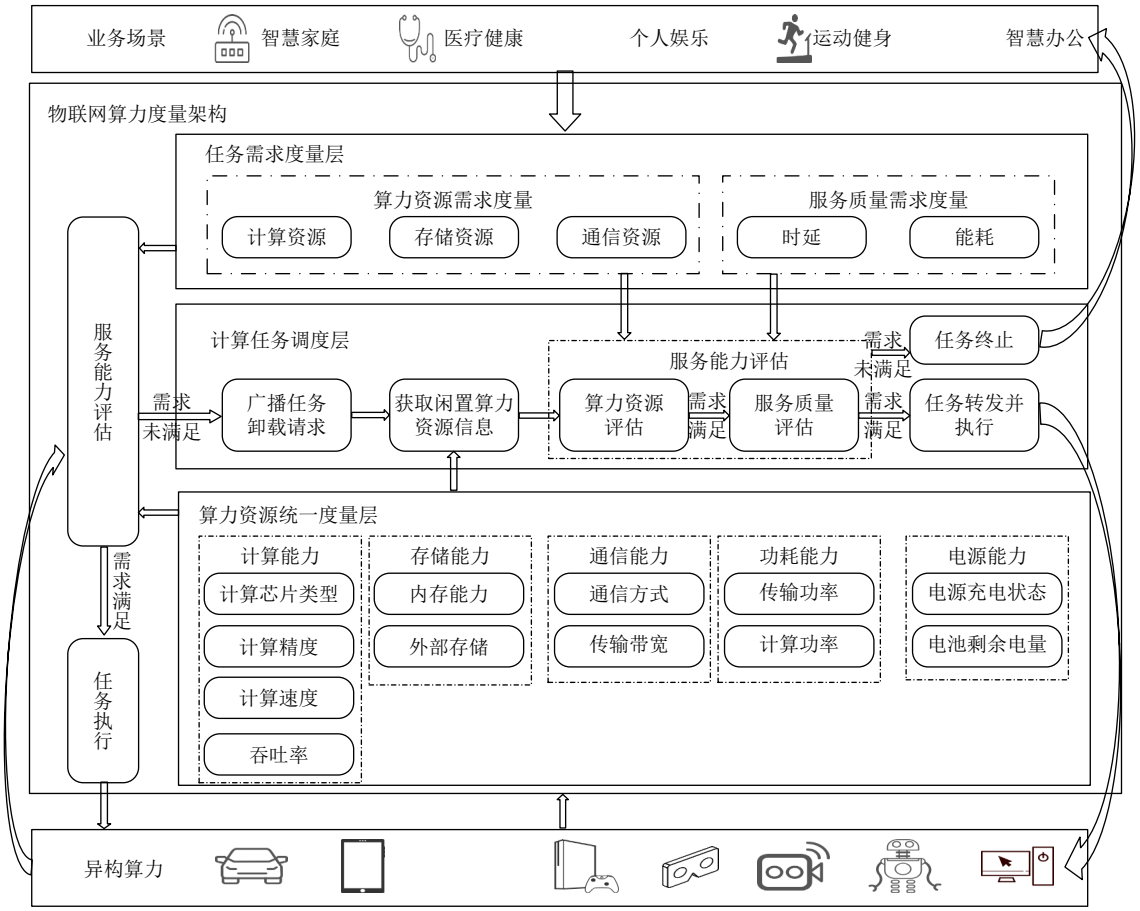

图1　面向物联网端侧设备的算力度量架构

2.1.2　任务需求度量层

当前的任务调度方法主要侧重于业务对计算资源的需求，特别是在保障服务质量以满足任务时延要求方面。然而，这些方法未能全面考虑算力资源与服务质量需求的综合关系，并忽视了物联网业务的能耗敏感特性。为了更有效地实现业务需求与算力资源的精确匹配，本文在任务需求度量层进行了创新性的改进。此层不仅为业务场景中的计算任务提供了详尽的算力资源需求度量指标，还特别增加了时延和能耗两个关键维度的服务质量需求度量。这种全面的度量方法不仅考虑了任务的计算需求，还充分考虑了物联网业务在时延和能耗方面的特殊需求。纳入这些多维度的服务质量指标，能够为后续算力资源的匹配提供更准确的判断依据，确保服务质量能够满足物联网业务在时效性和能效性上的双重需求。这一改进显著提高了任务调度的效率和准确性，同时确保了业务运行的经济性和可持续性。

2.1.3　计算任务调度层

在当前的中心化任务调度模式下，中心节点需要处理应用场景中的所有计算任务，这对中心节点的算力资源提出了较高的要求，无法适用于资源受限的物联网端侧设备。针对这一问题，本文提出了一种创新的分布式任务调度方法。在这种方法中，物联网端侧设备自主负责任务调度，仅须处理自身的任务，而不需要考虑其他设备的任务调度，大幅降低了调度的资源消耗。具体来说，在物联网的各类业务场景中，对于计算任务，任务需求度量层已完成了算力资源需求及服务质量需求的度量。同时，所有的物联网异构端

侧设备也经过了算力资源的统一度量。对于需要承载计算任务的端侧设备，首先进行自我评估，判断自身的算力资源是否能够满足计算任务的需求。如果能够满足，设备将使用自身的算力资源完成任务执行；如果不能满足，设备将通过广播任务卸载请求的方式，获取周边闲置设备的算力资源度量结果。综合考虑自身的任务需求和闲置设备的算力资源，再通过服务能力评估，设备将实现计算需求与算力资源的有效匹配。在此过程中，如果存在闲置设备的算力资源能够满足计算任务的需求，并且符合服务质量要求，则任务将被转发至该闲置设备执行。反之，如果没有合适的资源，任务将终止，并将情况反馈给业务场景。这种分布式调度方法有效解决了传统中心化调度方法在物联网端侧设备应用中的局限性，提高了任务调度的灵活性和效率，同时减少了对中心节点的算力依赖，更符合物联网设备分布广泛、资源受限的特点。

2.2　算力资源统一度量方法

本文针对异构物联网端侧设备的算力资源实施了一种综合的统一度量方法，这种方法包括计算、存储、通信、功耗和电源5个关键维度，并细化为12项细粒度度量指标，为物联网算力资源的动态共享和有效管理提供了坚实的基础。

假设物联网场景中共有N个端侧算力设备，对任意一个物联网设备D_i（$1 \leqslant i \leqslant N$）的算力资源度量结果用$D_i^{\mathrm{R}}$表示，对于每个端侧设备$D_i$，其在计算、存储、通信、功耗和电源5个维度的度量结果分别用R_i^{C}、R_i^{S}、R_i^{M}、R_i^{P}和R_i^{B}表示。因此，针对任意的物联网端侧设备D_i的算力资源度量情况D_i^{R}可表示为：

$$D_i^{\mathrm{R}} = \left\{R_i^{\mathrm{C}}, R_i^{\mathrm{S}}, R_i^{\mathrm{M}}, R_i^{\mathrm{P}}, R_i^{\mathrm{B}}\right\} \tag{1}$$

这种方法能够精确地评估每个端侧设备在这5个维度上的算力资源，为物联网设备间算力资源的高效分配和调度提供了可靠的量化依据。

2.2.1　计算能力度量

物联网端侧设备的计算能力度量R_i^{C}主要用于评估设备所配备的计算芯片性能及设备的运算能力。包括计算芯片类型$\bar{C}_i$、支持的计算精度$\bar{P}_i$，以及在不同精度下的运算速率$\bar{R}_i$，这些信息通常由芯片供应商提供。至于设备的实际运算能力则通过特定的测试程序或评测方法来获取，例如，测量该设备在执行对应运算任务时的吞吐率$\bar{T}_i$。物联网设备的计算能力R_i^{C}可以用式（2）表示。

$$R_i^{\mathrm{C}} = \left\{\bar{C}_i, \bar{P}_i, \bar{R}_i, \bar{T}_i\right\} \tag{2}$$

随着物联网应用场景的不断丰富，当前物联网端侧设备呈现内核芯片异构化的趋势，除了通用计算外，高性能计算和智能计算的比重越来越大、类型越来越多[18]。与云中心和边缘节点通常采用高性能CPU和GPU芯片不同，物联网设备所搭载的芯片通常异构程度较高、计算资源相对受限，如集成了相对较弱的CPU、GPU和NPU内核的系统级芯片（system on chip，SoC）、半定制的现场可编程门阵列（field programmable gate array，FPGA）、定制化的专用集成电路（application specific integrated circuit，ASIC）等。计算芯片类型$\bar{C}_i$的可能取值如式（3）所示。

$$\bar{C}_i \in \{\text{CPU, GPU, NPU, FPGA, ASIC, } \cdots\} \tag{3}$$

物联网设备上搭载的计算芯片类型$\bar{C}_i$不同，能够支持的数据计算精度$\bar{P}_i$也会不同，甚至多种计算精度可选，以适应混合精度运算的需求。例如，整数型计算精度可能包括INT4、INT8和INT16，而浮点数计算精度可能包括FLOAT16、FLOAT32和FLOAT64等。物联网端侧设备支持的数据计算精度$\bar{P}_i$如式（4）所示。

$$\bar{P}_i \in \{\text{INT4, INT8, FLOAT16, FLOAT32, } \cdots\} \tag{4}$$

对于不同精度下的运算速率$\bar{R}_i$，现有的度量通常以每秒所执行的浮点运算次数（floating point operations per second，FLOPS）和每秒运算

次数（operations per second，OPS）为标准，缺乏对不同精度运算速率的细粒度区分。本文提出的运算速率度量结果$\bar{R}_i$需要与具体的运算精度相匹配，表示为式（5）。

$$\bar{R}_i \in \left\{ \bar{R}_i^{\mathrm{INT4}}, \bar{R}_i^{\mathrm{INT8}}, \bar{R}_i^{\mathrm{FLOAT16}}, \bar{R}_i^{\mathrm{FLOAT32}}, \cdots \right\} \quad (5)$$

对于任意精度P下的运算速率$\bar{R}_i^P$，由支持该精度的处理器核数Q、主频F和该处理器单个时钟周期的运算次数D决定，如式（6）所示。

$$\bar{R}_i^P = Q \times F \times D \quad (6)$$

考虑物联网设备通常用于特定的智能算法运算（如人脸识别、动作识别、语音识别等），并且芯片在实际应用中通常无法完全发挥其所有计算单元的最大效能，因此除了标称运算速率$\bar{R}_i$之外，还需要考虑实际的训练或推理速率，即算法运算的吞吐率$\bar{T}_i$。这些速率可以通过MLPerf等评测方法获取，例如，运行目标检测算法YoLo的速率$\bar{T}_i^{\mathrm{YoLo}}$或姿态识别算法OpenPose的处理速率$\bar{T}_i^{\mathrm{OpenPose}}$。算法运算的吞吐率$\bar{T}_i$可表示为：

$$\bar{T}_i \in \left\{ \bar{T}^{\mathrm{YoLo}}, \bar{T}_i^{\mathrm{OpenPose}}, \cdots \right\} \quad (7)$$

设备完成特定推理任务Task的吞吐率$\bar{T}_i^{\mathrm{Task}}$如式（8）所示，表示单位时间下完成该推理任务的次数。

$$\bar{T}_i^{\mathrm{Task}} = \frac{N^{\mathrm{Task}}}{T} \quad (8)$$

其中，N^{Task}表示设备花费T时长能完成推理任务Task的次数。

这种方法能够更准确地度量物联网设备的计算能力，从而为算力资源的高效分配和任务调度提供可靠的量化基础。

2.2.2 存储能力度量

物联网端侧设备的存储能力度量R_i^{S}是评估设备处理数据的关键能力，涉及外部存储能力$\bar{O}_i$（如硬盘）和内部存储能力$\bar{I}_i$（如内存）。存储能力通常使用KB、MB、GB等单位表示。端侧设备的存储能力S包括外部存储能力$\bar{O}_i$和内部存储能力$\bar{I}_i$，都由设备对应存储器中的最小存储单元s和存储单元的数量K共同决定，如式（9）所示。

$$S = K \times s \quad (9)$$

外部存储能力$\bar{O}_i$指设备可以访问的非内置存储资源，如外接硬盘或网络存储设施。这部分存储通常用于大量数据的长期存储，关键在于其存储容量和数据传输速率。内部存储能力$\bar{I}_i$则涉及设备内置的存储资源，如RAM和内置硬盘。内部存储的重点在于快速读/写能力和临时存储数据的容量，对于设备的即时处理能力至关重要。因此，物联网端侧设备的存储能力R_i^{S}的度量综合这两方面的能力，如式（10）所示。

$$R_i^{\mathrm{S}} = \left\{ \bar{O}_i, \bar{I}_i \right\} \quad (10)$$

这种方法能够全面了解物联网端侧设备的存储能力，为数据处理和存储优化提供关键信息。适当的存储能力评估也有助于合理规划设备资源，优化整体网络的性能和响应能力。

2.2.3 通信能力度量

物联网设备间的连接离不开无线通信技术，这种技术的发展不仅使得算力资源更加便捷地泛在扩展，还确保了数据流动性和用户的使用便利性。由于不同通信技术支持不同的传输协议和信号标准，因此，不支持同一通信技术的设备之间无法实现互联。此外，不同设备支持的传输带宽也存在差异。对于特定设备D_i的通信能力度量R_i^{M}，应包括该设备所能支持的通信技术$\bar{S}_i$及其支持的传输带宽$\bar{B}_i$的度量。这样，在执行计算任务卸载时，可以根据通信能力的度量情况选取支持对应通信技术的设备进行通信和算力调度。R_i^{M}的度量可表述为：

$$R_i^{\mathrm{M}} = \left\{ \bar{S}_i, \bar{B}_i \right\} \quad (11)$$

其中，$\bar{S}_i$代表设备支持的通信方式，可以包括多种通信技术，如4G、5G等移动通信技术，Wi-Fi、蓝牙、ZigBee等短距离通信技术，窄带物联网（narrow band Internet of things，NB-IoT）、超窄

带技术（SigFox）等低功耗广域网通信技术，以及毫米波、无源射频识别（radio frequency identification，RFID）、超宽带（ultra wide band，UWB）等，考虑功耗和成本，特定物联网设备通常只部署有限的通信模块以满足业务需求，导致物联网设备间通信方式的多样性。因此，$\bar{S}_i$的可能值如式（12）所示。

$$\bar{S}_i \in \{4\mathrm{G},\ 5\mathrm{G},\ \mathrm{Wi\text{-}Fi},\ \mathrm{NB\text{-}IoT},\ \cdots\} \tag{12}$$

本文中传输带宽$\bar{B}_i$是设备在特定通信方式下的最大传输速率，如式（13）所示，可以采用香农公式进行量化计算，单位为bit/s，如式（14）所示。

$$\bar{B}_i \in \left\{\bar{B}_i^{4\mathrm{G}},\ \bar{B}_i^{5\mathrm{G}},\ \bar{B}_i^{\mathrm{Wi\text{-}Fi}},\ \bar{B}_i^{\mathrm{NB\text{-}IoT}},\ \cdots\right\} \tag{13}$$

$$\bar{B}_i^{C} = W^{C} \times \log\left(1 + (S/N)^{C}\right) \tag{14}$$

其中，W^C为在通信方式为C时的信道带宽，$(S/N)^C$为通信环境中的信噪比。

这种方法能够准确评估物联网端侧设备在各种通信技术下的通信能力，为算力资源分配和计算任务的有效调度提供重要的参考信息。

2.2.4　功耗情况度量

物联网设备在进行数据通信和计算任务处理过程中都会产生能量消耗。为了评估物联网设备的能耗情况，本文将物联网设备的功耗情况R_i^{P}纳入度量指标体系中。考虑设备的功耗与其通信带宽、计算时长以及计算任务和调度策略紧密相关，本文主要关注设备本身的运行功率。这包括其通信过程中的发射功率$\bar{P}_i^{\mathrm{T}}$和在处理计算任务时的运算功率$\bar{P}_i^{\mathrm{C}}$。因此，设备的运行功率R_i^{P}如式（15）所示。

$$R_i^{\mathrm{P}} = \left\{\bar{P}_i^{\mathrm{T}}, \bar{P}_i^{\mathrm{C}}\right\} \tag{15}$$

这种方法可以更准确地评估物联网设备在进行关键操作时的能量消耗。这对于设计能效高、持续时间长的物联网系统至关重要，特别是针对资源受限的应用场景。考虑能源效率是物联网设备可持续运行的关键因素，对功耗情况的准确度量不仅有助于优化设备性能，也对降低系统整体的运行成本和环境影响具有重要意义。精细化的功耗度量能够为物联网设备设计更为高效和环保的能源管理策略。

2.2.5　电源能力度量

由于物联网设备的可移动性且部分物联网设备不具备持续充电的条件，物联网端侧设备通常对功耗十分敏感。因此，在进行任务卸载时，需要充分考虑算力设备的电源能力。本文采用R_i^{B}来表示物联网端侧设备的电源能力度量情况，涵盖了设备的电源充电状态度量指标$\bar{L}_i$和设备电源剩余电量度量指标$\bar{U}_i$，如式（16）所示。

$$R_i^{\mathrm{B}} = \left\{\bar{L}_i, \bar{U}_i\right\} \tag{16}$$

其中，$\bar{L}_i$表示设备是否正在充电。如果设备D_i正在充电（$\bar{L}_i$=True），则表明该设备不存在电量耗尽的风险。相反，如果设备不在充电状态（$\bar{L}_i$=False），则需要特别注意其剩余电量，以确保它能够支持当前任务的完成。另外，$\bar{U}_i$衡量的是设备的剩余电量，这对于预判设备能否完成更长时间的计算任务至关重要。在物联网环境中，经常会出现设备需要在电量有限的情况下完成任务，因此准确评估设备的电源剩余电量对于确保任务顺利执行非常关键。综上所述，物联网端侧设备的电源能力度量不仅涉及其电源的当前状态和剩余电量，而且还应考虑设备的具体使用场景和任务需求。这种综合度量能够更有效地评估和规划设备的电源使用，确保物联网设备在执行任务时的稳定性和可靠性。

本文提出了一种面向物联网端侧设备的全面算力度量方法，该方法涵盖了计算（R_i^{C}）、通信（R_i^{M}）、存储（R_i^{S}）、功耗（R_i^{P}）和电源（R_i^{B}）这5个关键维度。这种方法能够对物联网端侧设备的异构算力资源进行全面的度量评估。对于任何一个物联网端侧设备D_i，其算力资源的度量结果D_i^{R}如式（17）所示。

$$D_i^{\mathrm{R}}=\begin{cases}R_i^{\mathrm{C}}=\begin{cases}\bar{C}_i\in\{\mathrm{CPU},\ \mathrm{GPU},\ \mathrm{MCU},\ \mathrm{FPGA},\ \mathrm{ASIC},\ \cdots\} & \text{//计算芯片类型}\\ \bar{P}_i\in\{\mathrm{INT4},\ \mathrm{INT8},\ \mathrm{FLOAT16},\ \mathrm{FLOAT32},\ \cdots\} & \text{//计算精度}\\ \bar{R}_i\in\left\{\bar{R}_i^{\mathrm{INT4}},\ \bar{R}_i^{\mathrm{INT8}},\ \bar{R}_i^{\mathrm{FLOAT16}},\ \bar{R}_i^{\mathrm{FLOAT32}},\ \cdots\right\} & \text{//不同精度下的计算速度}\\ \bar{T}_i\in\left\{\bar{T}_i^{\mathrm{Yolo}},\ \bar{T}_i^{\mathrm{OpenPose}},\ \cdots\right\} & \text{//在特定模型下的吞吐率}\end{cases}\\ R_i^{\mathrm{M}}=\begin{cases}\bar{S}_i\in\{\mathrm{4G},\ \mathrm{5G},\ \mathrm{Wi\text{-}Fi},\ \mathrm{NB\text{-}IoT},\ \cdots\} & \text{//设备支持的通信方式}\\ \bar{B}_i\in\left\{\bar{B}_i^{\mathrm{4G}},\ \bar{B}_i^{\mathrm{5G}},\ \bar{B}_i^{\mathrm{Wi\text{-}Fi}},\ \bar{B}_i^{\mathrm{NB\text{-}IoT}},\ \cdots\right\} & \text{//设备在指定通信方式下的传输带宽}\end{cases}\\ R_i^{\mathrm{S}}=\begin{cases}\bar{O}_i & \text{//设备的外部存储大小}\\ \bar{I}_i & \text{//设备的缓存大小}\end{cases}\\ R_i^{\mathrm{P}}=\begin{cases}\bar{P}_i^{\mathrm{T}} & \text{//设备通信功率}\\ \bar{P}_i^{\mathrm{C}} & \text{//设备计算功率}\end{cases}\\ R_i^{\mathrm{B}}=\begin{cases}\bar{L}_i & \text{//设备电源充电情况}\\ \bar{U}_i & \text{//设备剩余电量}\end{cases}\end{cases}\tag{17}$$

本方法共包含12项具体的度量指标，这些细粒度的度量指标能够准确地评估物联网设备的服务能力，以适应各种类型的业务需求。这种度量方法的优势在于其能够为物联网设备的算力资源提供一个通用的、可共享的评估框架，有助于实现算力资源的高效共享和流通。在物联网环境中，算力资源的合理分配和利用对于提高整体网络的性能和响应能力至关重要。使用这种全面且细致的度量方法，可以更好地优化物联网端侧设备在各种应用场景中的运行效率，推动物联网技术的进一步发展和应用。

2.3 任务需求度量方法

如第2.2节所述，假设在任意物联网业务场景中有N个端侧算力设备，每个设备D_i（$1\leqslant i\leqslant N$）的算力资源度量结果用D_i^{R}表示。在大模型和高性能计算不断发展的背景下，部分物联网设备可能无法满足不断增长的算力和服务质量需求。此外，由于潮汐效应，许多物联网设备在空闲时段存在算力资源浪费的现象。因此，通过任务卸载机制，可以将资源受限的端侧设备的计算任务卸载至资源充足且处于空闲状态的设备上，以实现算力资源的共享。

对于资源受限的物联网端侧设备D_j（$1\leqslant j\leqslant N$），当其自身算力资源无法承载第l个计算任务$\mathrm{Task}_{j,l}$（$1\leqslant j\leqslant N, 1\leqslant l$）时，可以通过使用第2.4节所提出的任务调度方法来满足任务$\mathrm{Task}_{j,l}$的计算资源和服务质量需求，并且用算力资源充足的空闲设备D_i来完成该任务的计算。本文中，任务$\mathrm{Task}_{j,l}$的需求度量结果用$T_{j,l}^{\mathrm{R}}$表示，具体如式（18）所示。该度量包括算力资源需求和服务质量需求两个方面。其中，算力资源需求的度量涉及计算（$T_{j,l}^{\mathrm{C}}$）、存储（$T_{j,l}^{\mathrm{S}}$）和通信（$T_{j,l}^{\mathrm{M}}$）3个维度；而服务质量需求则包括时延（$T_{j,l}^{\mathrm{D}}$）和能耗（$T_{j,l}^{\mathrm{E}}$）两个维度。这种综合的度量方法不仅有助于精确评估任务的需求，也为资源受限的物联网设备提供了更高效的任务调度和资源分配方案。

$$T_{j,l}^{\mathrm{R}}=\left\{T_{j,l}^{\mathrm{C}},T_{j,l}^{\mathrm{S}},T_{j,l}^{\mathrm{M}},T_{j,l}^{\mathrm{D}},T_{j,l}^{\mathrm{E}}\right\}\tag{18}$$

在物联网环境中，对于计算任务$\mathrm{Task}_{j,l}$的需求度量是至关重要的，以确保任务能够在合适的设备上高效执行。具体到每个任务$\mathrm{Task}_{j,l}$，其综合需求度量包括多个关键方面，具体如下。

（1）计算资源需求$T_{j,l}^{\mathrm{C}}$：这一部分包括任务所需的计算芯片类型$C_{j,l}$、计算精度$P_{j,l}$、计算速度$R_{j,l}$、特定计算任务的吞吐速率$T_{j,l}$，及所需完成的计算次数$N_{j,l}$或处理的图像帧数$F_{j,l}$。例如，

对于执行YOLO算法的计算任务，其需求可以转化为YOLO算法的运算要求，即 $T_{j,l}^{\mathrm{YoLo}}$。

（2）存储资源需求 $T_{j,l}^{\mathrm{S}}$：包括内部存储需求 $I_{j,l}$ 和外部存储需求 $O_{j,l}$。内部存储 $I_{j,l}$ 指的是计算任务运行时需要加载到内存中的数据量，而外部存储需求 $O_{j,l}$ 则是指计算任务所需处理数据的存储大小。

（3）通信需求 $T_{j,l}^{\mathrm{M}}$：涉及发起任务卸载请求的设备 D_j 支持的通信方式 $\bar{S}_j$ 和任务 $\mathrm{Task}_{j,l}$ 的数据传输要求对应的通信带宽 $B_{j,l}$。只有支持相同通信方式的闲置设备 D_i 才能响应 D_j 的任务卸载请求，即 $\bar{S}_j \cap \bar{S}_i \neq \varnothing$。

（4）时延需求 $T_{j,l}^{\mathrm{D}}$：指用户对于计算任务 $\mathrm{Task}_{j,l}$ 能够接受的时延阈值。

（5）能耗需求 $T_{j,l}^{\mathrm{E}}$：表示用户对于计算任务 $\mathrm{Task}_{j,l}$ 的能耗阈值。

综上所述，对于计算任务 $\mathrm{Task}_{j,l}$ 的需求度量可综合以上各个方面，如式（19）所示。

$$T_{j,l}^{\mathrm{R}}=\begin{cases} T_{j,l}^{\mathrm{C}}=\begin{cases} C_{j,l} \ \text{//计算芯片类型} \\ P_{j,l} \ \text{//计算精度} \\ T_{j,l} \text{或} R_{j,l} \ \text{//计算速率或特定模型的计算吞吐率} \\ N_{j,l} \text{或} F_{j,l} \ \text{//计算次数或需计算的图片帧数} \end{cases} \\ T_{j,l}^{\mathrm{S}}=\begin{cases} I_{j,l} \ \text{//内部缓存大小} \\ O_{j,l} \ \text{//外部存储大小} \end{cases} \\ T_{j,l}^{\mathrm{M}}=\begin{cases} \bar{S}_j \ \text{//任务请求设备支持的通信方式} \\ B_{j,l} \ \text{//任务请求设备的通信带宽} \end{cases} \\ T_{j,l}^{\mathrm{D}} \ \text{//完成任务的时延阈值} \\ T_{j,l}^{\mathrm{E}} \ \text{//完成任务的能耗阈值} \end{cases} \tag{19}$$

这种全面细致的任务需求度量方法为物联网端侧设备间的任务分配和资源利用奠定了坚实的基础。精确评估每个任务的具体需求，能够确保算力资源在物联网环境中得到最优的分配和利用，进而提高整个网络的性能和响应速度。

2.4　一种分布式任务调度方法

在物联网业务场景中，如何将物联网业务场景中的计算任务根据其需求度量情况调度至最匹配的物联网端侧算力节点，是提升物联网算力的共享能力和计算任务服务质量的关键。本文提出了一种分布式的任务调度方法，旨在优化物联网的计算资源分配和提高服务质量。该分布式任务调度流程如图2所示。当物联网端侧设备经过服务能力评估后无法继续承载任务时，它将通过广播消息的方式向周边设备请求任务卸载，并根据周边闲置设备的算力资源度量结果，结合当前设备的需求，进行算力资源评估和服务质量评估，以实现有效的计算需求和算力资源匹配。这一过程包括选择合适的闲置设备和任务的调度下发，确保用户能够持续享受高质量的物联网算力服务。如果所有可通信的物联网端侧设备中都找不到合适的设备来承载该计算任务，则该任务将被中止。整个任务调度过程主要包括以下5个步骤。

步骤1　广播任务卸载请求

在物联网业务场景中，算力和服务质量的需求持续升高。因此，部分端侧设备可能无法满足特定计算任务 $\mathrm{Task}_{j,l}$ 的算力或服务质量需求。这些设备 D_j 可以向周边能够通信的所有物联网端侧设备广播自己的任务卸载请求，并等待响应。为降低通信时延，广播仅限于一跳。本方法不考虑任务拆分执行的方案，因此单个计算任务需要由单一的计算设备独立完成。

步骤2　获取闲置算力度量信息

为了响应物联网端侧设备 D_i 的任务卸载请求，所有物联网端侧设备 D_i 的资源使用状态由 U_i 表示。当 $U_i=1$，表示设备 D_i 的资源已被占用，当前无法供其他设备使用；若 $U_i=0$，则表示设备 D_i 的资源处于空闲状态，可供其他设备共享。当设备 D_i 接收到来自设备 D_j 的任务卸载请求，且其资源处于空闲状态（$U_i=0$）时，设备 D_i 应将自己的算力资源度量结果 D_i^{R} 发送给请求设备 D_j；如果设备 D_i 的资源正在被使用（$U_i=1$），则不对设备 D_i 的任务卸载请求做出响应。

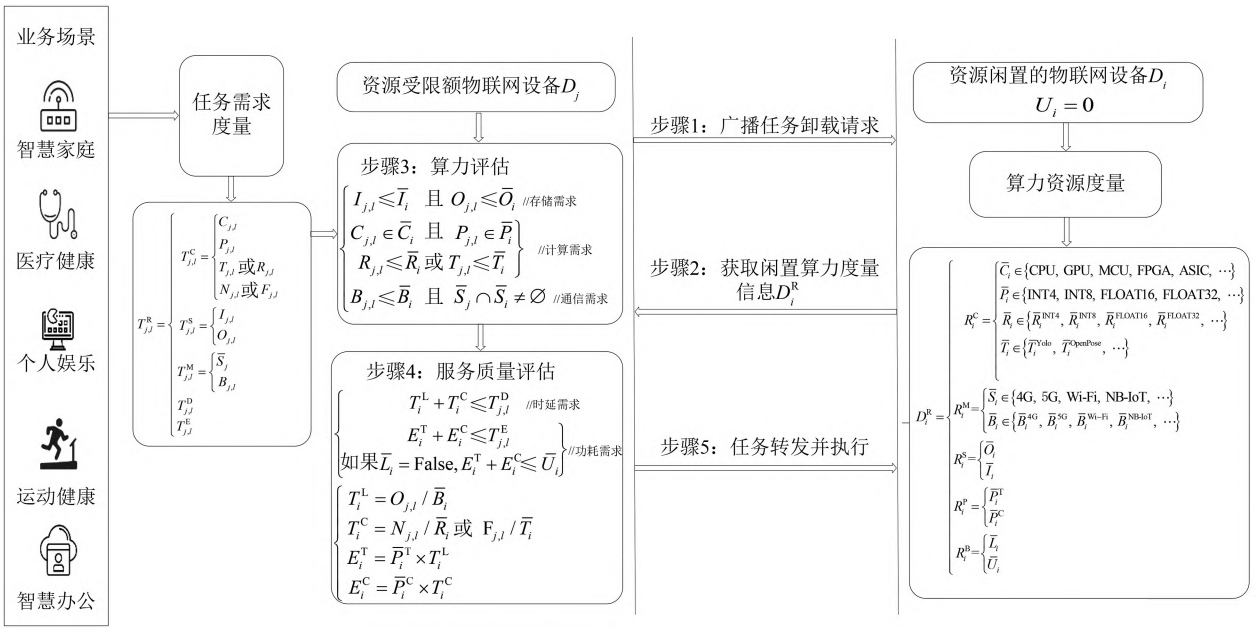

图 2　分布式任务调度流程

步骤 3　算力评估

算力评估的主要目的是判断端侧设备D_i的算力资源是否满足计算任务$\mathrm{Task}_{j,l}$的需求$T_{j,l}^{\mathrm{R}}$，其中包括计算、存储和通信等资源，该评估过程需要满足以下约束条件。

（1）存储需求：设备D_i的内部存储空间$\bar{I}_i$和外部存储空间$\bar{O}_i$必须分别大于或等于任务$\mathrm{Task}_{j,l}$的内部存储需求$I_{j,l}$和外部存储需求$O_{j,l}$。

（2）计算需求：设备D_i的计算能力$\bar{C}_i$和计算精度$\bar{P}_i$应满足任务$\mathrm{Task}_{j,l}$的要求。

（3）通信需求：设备D_i的通信带宽$\bar{B}_i$应满足任务$\mathrm{Task}_{j,l}$的通信需求，同时设备D_i和D_j应支持相同的通信方式。

综合以上条件，算力评估的过程可以表示为满足式（20）的约束。

$$\begin{cases} I_{j,l}\leqslant\bar{I}_i \text{ 且 } O_{j,l}\leqslant\bar{O}_i \text{ //存储需求} \\ \left.\begin{aligned} C_{j,l}\in\bar{C}_i \text{ 且 } P_{j,l}\in\bar{P}_i \\ R_{j,l}\leqslant\bar{R}_i \text{ 或 } T_{j,l}\leqslant\bar{T}_i \end{aligned}\right\} \text{ //计算需求} \\ B_{j,l}\leqslant\bar{B}_i \text{ 且 } \bar{S}_j\cap\bar{S}_i\neq\varnothing \text{ //通信需求} \end{cases} \tag{20}$$

这一评估流程可以确保物联网端侧设备在任务卸载过程中选择适合的设备完成计算任务，从而优化整个物联网系统的性能和响应速度。

步骤 4　服务质量评估

在本文中，服务质量评估是根据端侧设备D_j需要卸载的计算任务$\mathrm{Task}_{j,l}$的服务质量度量执行的。该度量包括时延和能耗需求，用于判断端侧设备D_i是否能够承载计算任务$\mathrm{Task}_{j,l}$。计算任务$\mathrm{Task}_{j,l}$在传输过程需要传输的数据量为$O_{j,l}$，计算过程需要完成的计算次数或处理的图片帧数分别为$N_{j,l}$和$F_{j,l}$。卸载到设备D_i的计算任务$\mathrm{Task}_{j,l}$将产生传输时延T_i^{L}和计算时延T_i^{C}，以及相应的传输能耗E_i^{T}和计算能耗E_i^{C}。时延及能耗的计算可表示为：

$$\begin{cases} T_i^{\mathrm{L}}=O_{j,l}/\bar{B}_i \\ T_i^{\mathrm{C}}=N_{j,l}/\bar{R}_i \text{ 或 } F_{j,l}/\bar{T}_i \\ E_i^{\mathrm{T}}=\bar{P}_i^{\mathrm{T}}\times T_i^{\mathrm{L}} \\ E_i^{\mathrm{C}}=\bar{P}_i^{\mathrm{C}}\times T_i^{\mathrm{C}} \end{cases} \tag{21}$$

为了满足计算任务$\mathrm{Task}_{j,l}$的服务质量需求，响应任务的空闲端侧设备D_i必须满足以下时延和能耗约束条件：数据传输时延T_i^{L}和计算时延T_i^{C}之和应小于任务的需求时延$T_{j,l}^{\mathrm{D}}$。若设备D_i为能耗敏感型设备（$\bar{L}_i=\mathrm{True}$），则传输能耗E_i^{T}和计

算能耗E_i^{C}之和不仅要小于任务的需求$T_{j,l}^{\mathrm{E}}$，还需要保证设备D_i的剩余电量足够支撑完成任务。因此，根据式（22）约束选取能够满足计算任务$\mathrm{Task}_{j,l}$服务质量需求的端侧设备D_i。

$$\begin{cases} T_i^{\mathrm{L}}+T_i^{\mathrm{C}}\leqslant T_{j,l}^{\mathrm{D}} & \text{//时延需求} \\ E_i^{\mathrm{T}}+E_i^{\mathrm{C}}\leqslant T_{j,l}^{\mathrm{E}} & \\ \text{如果 } \bar{L}_i=\mathrm{False}, E_i^{\mathrm{T}}+E_i^{\mathrm{C}}\leqslant \bar{U}_i & \text{//功耗需求} \end{cases} \tag{22}$$

这种服务质量评估可以确保物联网环境中的任务卸载和算力资源分配不仅基于算力的匹配，而且符合时延和能耗的关键服务质量标准，从而优化整个物联网系统的性能和用户体验。

步骤5　任务转发并执行

在物联网环境中，任务转发和执行是分布式任务调度方法的最后且至关重要的一步。算力资源受限的端侧设备D_j经过步骤1和步骤2获取到周边闲置设备D_i的算力资源度量结果D_i^{R}，然后根据闲置设备的资源度量结果D_i^{R}和设备D_j需要卸载的计算任务$\mathrm{Task}_{j,l}$的需求度量情况$T_{j,l}^{\mathrm{R}}$来完成步骤3的算力评估（式（20））和步骤4的服务质量评估（式（22））。后续具体的任务转发情况及执行方法如下。

（1）若当前没有闲置设备通过算力评估和服务质量评估，则任务$\mathrm{Task}_{j,l}$将被终止，设备D_j将情况反馈给业务场景。

（2）若仅有1个闲置设备D_i通过算力评估和服务质量评估，则设备D_j将任务$\mathrm{Task}_{j,l}$转发给设备D_i，设备D_i将利用其算力资源来完成计算任务$\mathrm{Task}_{j,l}$。

（3）若有多个闲置设备D_i都能通过任务$\mathrm{Task}_{j,l}$的算力评估和服务质量评估，则进一步根据时延最短或能耗最小的策略选取最合适的闲置设备进行任务卸载，这样不仅保证了任务的快速执行，还优化了整体能耗，符合物联网设备的效率和环保要求。

• 对于时延敏感型任务，选取总时延（传输时延加计算时延）最短的设备，即：

$$\min\left(T_i^{\mathrm{L}}+T_i^{\mathrm{C}}\right), 1\leqslant i\leqslant N \tag{23}$$

• 对于功耗敏感型任务，选取总功耗（传输功耗加计算功耗）最小的设备，即：

$$\min\left(E_i^{\mathrm{T}}+E_i^{\mathrm{C}}\right), 1\leqslant i\leqslant N \tag{24}$$

这种方法可以确保物联网端侧设备的计算任务在最适合的设备上得到执行，最大化地利用了可用的算力资源，同时也优化了任务执行的时效性和能效性。这对于保持物联网系统的高效运作和提高用户体验至关重要。这种策略也体现了对物联网设备资源限制和服务质量需求的深入理解，有助于提升整个物联网环境的智能化和自适应能力。

3　智慧家庭设备算力度量和任务调度示例

物联网通过互联网连接各种物理设备和对象，使它们能够相互通信和协作，从而实现数据收集、分析和控制的能力，为人们的生活和工作带来了极大的便利。以智慧家庭场景为例，用户可以通过手机或语音设备来控制家庭电器，如灯光、空调、电视等，此外，还可以通过摄像机、智能门锁实现家庭安防监控、环境监测等多种功能，极大地提高了生活质量。本文选取智慧家庭应用场景为应用示例，评估所提出的算力度量架构及其在实际应用中的有效性，智慧家庭场景示例如图3所示。

3.1　场景说明

智慧家庭通过物联网设备可以感知用户在家中做的任何事情，随时能够通过智能化的功能，让生活更加便捷和安全。同时，针对用户的及时性需求，提供高效、低时延的服务，并且可以通过智能控制优化设备的使用，降低能源消耗，智慧家庭已经成为现代家居的新趋势，具有智能化、及时、安全和节能环保等特点。随着智慧家庭智能化水平的不断提升，各类智能算法的不断更迭对物联网设备的算力提出了更高要求，智慧

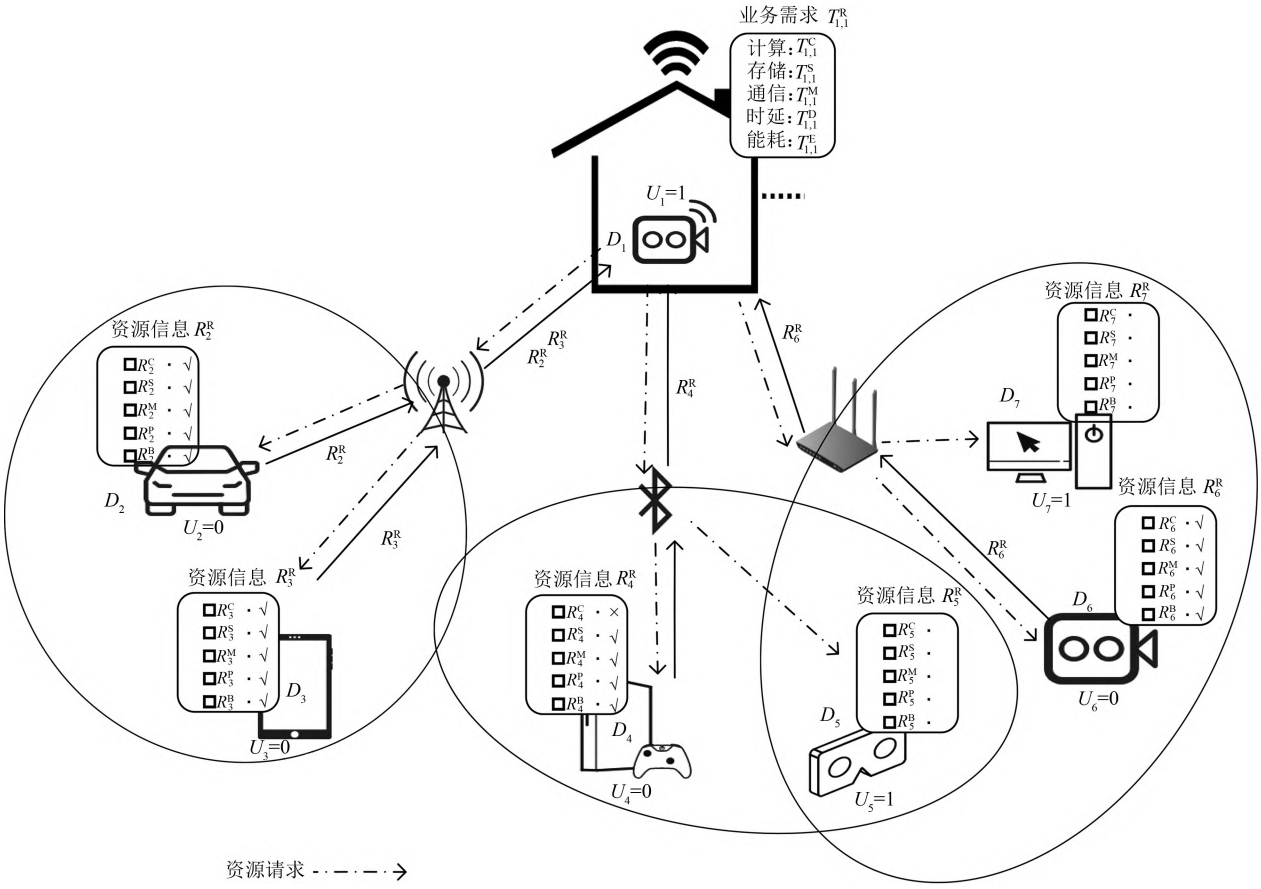

图3　智慧家庭场景示例

家庭中部分物联网设备会由于智能算法的升级而出现算力资源不足的情况，需要和周边设备交互进行任务的卸载。

图3描述的智慧家庭场景中，存在着多种具备计算能力的物联网设备，如摄像机、智能音响、游戏机、机顶盒等。这些设备能够通过不同的通信方式与周边设备进行交互。如摄像机D_1支持4G、蓝牙和Wi-Fi这3种通信方式；汽车D_2和手机D_3支持4G的通信方式和摄像机D_1通信；游戏机D_4和眼镜D_5则支持蓝牙方式和摄像机D_1通信；眼镜D_5、摄像机D_6和计算机D_7同样支持蓝牙通信。所有设备根据第2.2节提出的算力度量方法对算力资源进行全面评估，评估后得到度量结果为$M=\left\{D_1^{\mathrm{R}},D_2^{\mathrm{R}},\cdots,D_i^{\mathrm{R}}\right\}$，$i\in\{1,2,\cdots,7\}$。当摄像机$D_1$的算力资源不足以满足模型升级后的人脸识别任务$T_{1,1}$时，任务$T_{1,1}$的任务需求度量结果为$T_{1,1}^{\mathrm{R}}$。参考文献[19]的终端芯片数据及智慧家庭中常用设备的相关参数，本文智慧家庭任务需求度量情况见表1，智慧家庭设备算力度量情况见表2。本文提出的任务调度策略可以实现智慧家庭中的算力资源共享，从而有效地调度任务，同时满足任务的计算资源和服务质量需求。该场景下任务调度的具体实现流程在第3.2节详细说明。

由该示例可以看到，在物联网环境中，算力度量和任务调度的重要性为不仅提升了设备利用率和服务质量，还提高了系统的灵活性和效率，从而为用户带来更加丰富和便利的应用体验。

3.2　任务调度流程

智慧家庭场景中的任务调度流程可分为以下几个步骤。

表1　智慧家庭任务需求度量情况

需求大类	计算资源需求				存储需求		通信需求		时延/s	能耗/J
需求指标	计算芯片类型	计算精度	计算吞吐率/（f·s^{-1}）	图片帧数	内部存储/MB	外部存储/MB	通信方式	传输带宽/（Mbit·s^{-1}）		
任务$T_{1,1}$	GPU	INT8	20	1 000	2	200	4G、蓝牙、Wi-Fi	2	50	500

表2　智慧家庭设备算力度量情况

度量大类	度量指标	摄像头D_1	汽车D_2	手机D_3	游戏机D_4	眼镜D_5	摄像机D_6	计算机D_7
计算能力	计算芯片类型	CPU GPU	CPU GPU	CPU GPU	CPU GPU	CPU GPU	CPU GPU	CPU GPU
	支持的计算精度	INT8	INT8	INT8	INT8	INT8	INT8	INT8
	运算速率/TOPS	5	275	16	2	5	6	25
	吞吐率/（f·s^{-1}）	18	100	50	15	25	30	80
存储能力	内部存储/GB	2	32	8	16	8	2	8
	外部存储/GB	128	64	512	128	32	256	512
通信能力	通信方式	4G、蓝牙、Wi-Fi	4G	4G	蓝牙	蓝牙	Wi-Fi	Wi-Fi
	传输带宽/（Mbit·s^{-1}）	40	100	100	2	2	200	200
功耗情况	传输功率/W	0.100	0.500	0.400	0.036	0.036	0.064	0.064
	计算功率/W	15	45	8	20	10	10	120
电源能力	电源充电状态	True	False	False	False	False	True	True
	电池剩余电量/（kW·h）		20.0	12.0	3.6	12.0		

（1）任务广播请求：摄像机D_1向周边所有通信可达的设备广播任务卸载请求。在本例中，汽车D_2、手机D_3、游戏机D_4、眼镜D_5、摄像机D_6和计算机D_7都能收到摄像机D_1的请求。

（2）响应任务请求：接收到任务卸载请求的设备会根据自身的算力资源状态（空闲或占用）进行响应。设备的资源状态由U_i表示，$U_i=1$代表设备资源占用且不响应该请求（如眼镜D_5和计算机D_7），而$U_i=0$表示设备空闲并响应请求。在本示例中，汽车D_2、手机D_3、游戏机D_4、摄像机D_6均处于空闲状态，因此将自身的算力资源度量结果D_i^R（$i\in\{2,3,4,6\}$）反馈给摄像机D_1。

$$\begin{cases} I_{1,1}\leqslant\bar{I}_i \text{ 且 } O_{1,1}\leqslant\bar{O}_i & \text{//存储需求} \\ \left.\begin{array}{l} C_{1,1}\in\bar{C}_i \text{ 且 } P_{1,1}\in\bar{P}_i \\ T_{1,1}\leqslant\bar{T}_i \end{array}\right\} & \text{//计算需求} \\ B_{1,1}\leqslant\bar{B}_i & \text{//通信需求} \end{cases} \tag{25}$$

（3）服务能力评估：摄像机D_1根据任务$T_{1,1}$的任务需求度量结果为$T_{1,1}^{\mathrm{R}}$和收到的闲置算力资源度量结果D_i^R（$i\in\{2,3,4,6\}$）进行服务能力评估，包括算力资源评估和服务质量评估。首先评估这些资源是否能满足任务的计算需求，需要满足式（25）中的约束条件。经过评估，汽车D_2、手机D_3和摄像机D_6能满足要求，但游戏机D_4处理人脸识别任务$T_{1,1}$的吞吐率为15 FPS，无法满足任务$T_{1,1}$的20 FPS的计算吞吐率需求。之后对汽车D_2、手机D_3、摄像机D_6的算力资源D_i^R（$i\in\{2,3,6\}$）展开服务能力评估，包括时延和能耗，该评估需要满足式（26）的约束。根据式（26）中时延和能耗计算方式对汽车D_2、手机D_3、摄像机D_6的时延和能耗情况仿真结果如图4所示，对比表1中任务$T_{1,1}$的时延和能耗要求阈值，显示所有设备均满足服务能力需求。

$$\begin{cases} T_i^{\mathrm{L}} + T_i^{\mathrm{C}} \leqslant T_{1,1}^{\mathrm{D}} \ \text{//时延需求} \\ E_i^{\mathrm{T}} + E_i^{\mathrm{C}} \leqslant T_{1,1}^{\mathrm{E}} \\ \text{如果 } \bar{L}_i = \mathrm{False}, E_i^{\mathrm{T}} + E_i^{\mathrm{C}} \leqslant \bar{U}_i \end{cases} \text{//能耗需求}$$

s.t.

$$\begin{cases} T_i^{\mathrm{L}} = O_{1,1} / \bar{B}_i \\ T_i^{\mathrm{C}} = \mathrm{F}_{1,1} / \bar{T}_i \\ E_i^{\mathrm{T}} = \bar{P}_i^{T} \times T_i^{\mathrm{L}} \\ E_i^{\mathrm{C}} = \bar{P}_i^{C} \times T_i^{\mathrm{C}} \end{cases} \tag{26}$$

通过上述流程，智慧家庭中的算力资源可以被有效利用，以满足各种计算任务的需求。这种分布式任务调度方法不仅提高了资源利用效率，还确保了服务质量，从而在智慧家庭场景中实现了高效且可靠的智能服务。

3.3 示例结果分析

在本研究中，任务卸载决策仅考虑单个最适合的资源闲置设备。当存在多个设备可作为任务卸载的候选时，需要对这些设备的时延和能耗情况进行仿真计算，之后按最佳能效和最短时延来进行最终的卸载设备选取，复用其闲置资源来处理计算任务，以确保计算任务的高效和节能执行。本文智慧家庭中满足任务卸载的设备时延和能耗情况仿真结果如图4所示。

（1）针对特定类型任务的选取策略：若计算任务有说明是能耗敏感型任务或时延敏感型任务，则选取评估中能耗最小或时延最低的候选设备进行任务卸载。

• 对于时延敏感型任务，选取总时延（传输时延加计算时延）最短的设备，图4中汽车D_2的时延最低，即选取汽车D_2进行资源复用，完成任务$T_{1,1}$的计算。

• 对于功耗敏感型任务，选取总能耗（传输能耗加计算能耗）最小的设备，图4中手机D_3的能耗最低，即选取手机D_3进行资源复用，完成任务$T_{1,1}$的计算。

（2）无特别要求任务的选择标准：若计算任务没有特别要求，则综合考虑能效最优和时延最低两个原则进行设备选择。

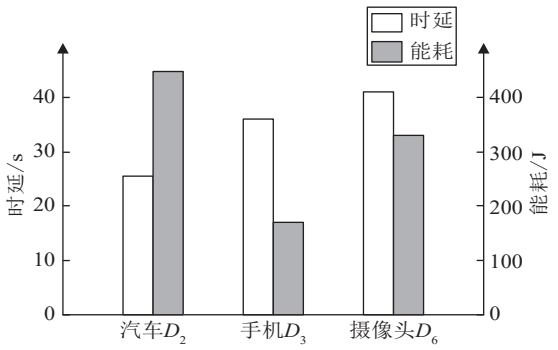

图4 时延和能耗情况仿真结果

• 对于能效最优，优先考虑连接电源的设备，以避免非连接电源的设备因承载额外任务而电量过低。

• 对于时延最低，在满足能效要求的前提下，选取完成任务时延最短的设备。

根据以上标准，在示例中，汽车D_2和手机D_3属于移动设备，没有连接电源，属于能耗敏感的端侧设备，被排除在外，最终选取连接电源且满足卸载任务$T_{1,1}$的算力资源和服务质量需求的摄像机D_6来进行资源复用，完成任务$T_{1,1}$的计算。

此示例分析显示了在智慧家庭场景下，如何根据物联网端侧设备的具体情况和任务需求，灵活地选择合适的设备进行任务卸载。这种方法不仅提高了任务执行的效率和效果，也最大限度地节约了能源消耗，符合智慧家庭对于资源优化和环境保护的普遍需求。

4 综论与未来研究方向

本文所提出的物联网算力度量架构，可以实现智慧家庭场景算力资源的高效共享。通过这一架构，可以在保证计算任务需求的同时，为用户提供低时延、低能耗的高质量物联网应用服务。但是本文所提出的方法仍存在一些局限，未来的研究将聚焦于以下3个方面。

（1）算力资源动态变化的处理：目前的算力度量方法和任务调度策略对于算力资源动态变化

的情况处理相对简单，仅考虑了设备的剩余电量和算力资源的占用状态，并假定在整个调度过程中算力资源保持不变。未来的研究需要开发更为动态的算力资源状态更新策略，以适应实际应用中算力资源的变化。

（2）资源共享的多样化：当前的任务调度方法在资源共享方面较为单一，仅考虑将整个任务调度至单一的资源空闲设备上进行处理，且一个设备在同一时间内不能处理多个任务。未来的工作需要探索任务拆分和跨多个设备调度的策略，以及单个设备处理多任务的可能性，从而提高资源利用效率和任务处理的灵活性。

（3）复杂调度场景的处理：本文的方法在处理物联网应用场景中的调度问题时相对简化，没有考虑多任务并发执行的情况。当物联网应用场景中存在多个任务需要同时执行时，如何有效协调多任务和多设备之间的关系，以实现全局资源共享和优化任务调度，是未来研究的重要课题。

综上所述，虽然本文提出的方法在智慧家庭场景中有效，但仍需进一步的优化和发展，以更好地适应物联网环境的复杂性和动态性。未来的研究将致力于提升算力资源管理的灵活性和智能性，从而为物联网用户提供更加高效、可靠的服务。

5　结束语

本文提出了一种面向物联网端侧设备的新型算力度量架构。这一架构为异构物联网端侧算力资源提供了一个全面的度量体系，涵盖了计算、存储、通信、功耗和电源这 5 个关键维度。此外，本文还特别针对物联网业务场景中的计算任务，分别对算力资源需求和服务质量需求进行了详细的度量。在此基础上，本文进一步提出了一种分布式的任务调度策略，实现了物联网算力资源与业务场景需求之间的智能匹配。通过在智慧家庭场景中的应用示例，本文方法已被证明能够有效地支持物联网算力资源的共享和任务调度，从而提升物联网计算效率并减少能源消耗。

下一步的研究方向将聚焦于更加动态和复杂的物联网应用场景，特别是针对算力资源的动态变化和计算任务的并发执行，计划进一步完善和丰富我们的算力度量架构。同时，还将探索更加全面的服务质量最优化的任务调度策略。这些研究不仅将提高物联网系统的灵活性和响应能力，而且还将为用户带来更高效、可靠的物联网服务体验。

参考文献：

[1] 杨光, 王玉申, 姚洁, 等. 算力时代下的算力服务需求研究[J]. 中国新通信, 2023, 25(1): 39-41.
YANG G, WANG Y S, YAO J, et al. Research on computing service demand in computing era[J]. China New Telecommunications, 2023, 25(1): 39-41.

[2] 中国移动通信集团有限公司. 算力网络白皮书[R]. 2021.
China Mobile Communications Group Co., Ltd. Computing force network technology white paper[R]. 2021.

[3] 姚惠娟, 陆璐, 段晓东. 算力感知网络架构与关键技术[J]. 中兴通讯技术, 2021, 27(3): 7-11.
YAO H J, LU L, DUAN X D. Architecture and key technologies for computing-aware networking[J]. ZTE Technology Journal, 2021, 27(3): 7-11.

[4] 何涛, 杨振东, 曹畅, 等. 算力网络发展中的若干关键技术问题分析[J]. 电信科学, 2022, 38(6): 62-70.
HE T, YANG Z D, CAO C, et al. Analysis of some key technical problems in the development of computing power network[J]. Telecommunications Science, 2022, 38(6): 62-70.

[5] 乔楚. 算力度量与算网资源调度思路分析[J]. 通信技术, 2022, 55(9): 1165-1170.
QIAO C. Analysis of the computing power measurement and resource scheduling on CPN[J]. Communications Technology, 2022, 55(9): 1165-1170.

[6] ITU-T. Computing power network-framework and architecture: Y.2501[S]. 2021.

[7] CCSA. 算力网络总体技术要求[R]. 2021.
CCSA. Computing power network overall technical requirements[R]. 2021.

[8] CCSA. 面向算网融合的算力度量与算力建模研究[R]. 2021.
CCSA. Research on computing power measurement and modeling for computing network fusion[R]. 2021

[9] 李建飞, 曹畅, 李奥, 等. 算力网络中面向业务体验的算力建

模[J]. 中兴通讯技术, 2020, 26(5): 34-38, 52.
LI J F, CAO C, LI A, et al. Computing power modeling for business experience in computing power network[J]. ZTE Technology Journal, 2020, 26(5): 34-38, 52.
[10] 柴若楠, 郜帅, 兰江雨, 等. 算力网络中高效算力资源度量方法[J]. 计算机研究与发展, 2023, 60(4): 763-771.
CHAI R N, GAO S, LAN J Y, et al. Efficient computing resource metric method in computing-first network[J]. Journal of Computer Research and Development, 2023, 60(4): 763-771.
[11] 郭亮, 吴美希, 王峰, 等. 数据中心算力评估: 现状与机遇[J]. 信息通信技术与政策, 2021(2): 79-86.
GUO L, WU M X, WANG F, et al. Research on evaluation of computing power and efficiency in data center: status and opportunities[J]. Information and Communications Technology and Policy, 2021(2): 79-86.
[12] 姜海洋, 李勇. 端边云场景下的算力度量方法[J]. 电信工程技术与标准化, 2023, 36(7): 79-83.
JIANG H Y, LI Y. Explore the correlation method between computing power measurement and service deployment in the device-edge-cloud collaboration scenario[J]. Telecom Engineering Technics and Standardization, 2023, 36(7): 79-83.
[13] 杜宗鹏, 李志强, 陆璐. 算力网络四面三级算力度量技术体系[J]. 中兴通讯技术, 2023, 29(4): 8-13.
DU Z P, LI Z Q, LU L. Three-level and four-aspect computing measurement system in computing force network[J]. ZTE Technology Journal, 2023, 29(4): 8-13.
[14] LI J C, LYU H, LEI B, et al. A computing power resource modeling approach for computing power network[C]//Proceedings of the 2022 International Conference on Computer Communications and Networks (ICCCN). Piscataway: IEEE Press, 2022: 1-2.
[15] 夏天豪, 夏长清, 潘昊, 等. 基于强化学习的算力资源度量方法[J]. 燕山大学学报, 2023, 47(3): 246-254.
XIA T H, XIA C Q, PAN H, et al. Computational power resource measurement method based on reinforcement learning[J]. Journal of Yanshan University, 2023, 47(3): 246-254.
[16] 周舸帆, 雷波. 算力网络中基于算力标识的算力服务需求匹配[J]. 数据与计算发展前沿, 2022, 4(6): 20-28.
ZHOU G F, LEI B. Computing service demand matching based on computing power identification in computing power network[J]. Frontiers of Data & Computing, 2022, 4(6): 20-28.
[17] 庞冉, 易昕昕, 辛亮, 等. 算力网络路由调度技术研究[J]. 电信科学, 2023, 39(8): 149-156.
PANG R, YI X X, XIN L, et al. Research on routing scheduling technology of computing power network[J]. Telecommunications Science, 2023, 39(8): 149-156.
[18] 中国信息通信研究院. 中国算力发展指数白皮书[R]. 2021.
China Academy of Information and Communications Technology. White paper on China computing power development index[R]. 2021.
[19] 中国人工智能产业发展联盟. 中国人工智能产业发展联盟AI芯片技术选型目录[EB]. 2020.
Artificial Intelligence Industry Alliance. Artificial Intelligence Industry Alliance AI chip technology selection catalog[EB]. 2020.

[作者简介]

祝淑琼（1994– ），女，中国移动通信有限公司研究院助理工程师，主要研究方向为算力度量和边缘计算技术。

徐青青（1987– ），女，中国移动通信有限公司研究院工程师，主要研究方向为行业智能和新型计算技术。

李小涛（1987– ），男，博士，中国移动通信有限公司研究院高级工程师，主要研究方向为物联网语义和新型计算技术。

陈维（1960– ），男，中国移动通信有限公司研究院首席科学家，主要研究方向为机器智能和边缘计算。