面向多用户的目标导向通信与节能型边缘资源管理

Francesco Binucci[1],Paolo Banelli[1], *IEEE* 会员，Paolo Di Lorenzo[2], *IEEE* 高级会员，以及 Sergio Barbarossa[2],*IEEE* 会士[1] 佩鲁贾大学，意大利佩鲁贾 Via G. Duranti 93, 06125；[2] 罗马第一大学，意大利罗马 Via Eudossiana 18，00184；电子邮箱：{francesco.binucci, paolo.banelli}@unipg.it，{paolo.dilorenzo, sergio.barbarossa}@uniroma1.it

arXiv:2305.02137v1 [eess.SP] 3 May 2023

摘要—**边缘学习（EL）将计算资源推向5G/6G网络边缘，以协助请求低时延且注重能耗的智能服务的移动用户。在远程执行推理任务时，一个常见挑战是仅提取并传输对推理任务最重要的特征。从这一角度来看，EL可以与目标导向通信有效结合，其目标是在规定的准确率、时延和能耗约束下，仅传输**与执行推理任务相关的**信息。在本文中，我们考虑一个多用户/单服务器无线网络，其中用户可以机会性地决定自行执行推理任务，或者将数据卸载到边缘服务器进行远程处理。待传输的数据经过目标导向压缩阶段，该阶段采用卷积编码器实现，并与运行于边缘服务器侧的卷积解码器联合训练。借助李雅普诺夫优化，我们提出了一种方法，用于联合且动态地优化最合适编码/解码方案的选择，以及所有用户和边缘服务器之间计算资源与传输资源的分配。大量仿真验证了所提方法的有效性，并突出了能耗、时延和学习准确率之间的权衡。**

索引词—**边缘学习、目标导向通信、Lyapunov随机优化、深度学习。**

I. 引言

THE 第五/第六代移动通信的出现彻底改变了网络概念，使其从纯粹的通信基础设施转变为普适服务的关键赋能者，而这些服务高度依赖人工智能（AI）和机器学习（ML）。增强现实、自动驾驶、大规模物联网和关键任务应用都是典型示例 [1]。在这些场景中，服务时延和可靠性约束往往十分严格，因此有必要设计一个整体系统，对通信、计算、学习和控制进行联合管理，以实现可靠性、能效和可持续性。

通过适当的 AI/ML 技术实时处理海量数据的需求，促使研究人员

本研究由 MIUR 根据 PRIN 2017 Liquid Edge 合同提供支持。P. Di Lorenzo 和 S. Barbarossa 的工作由 SNS-JU-2022 项目 ADROIT6G 资助，协议编号为 101095363。本研究还部分获得欧盟通过 NextGenerationEU 意大利国家复苏与韧性计划（NRRP）下的“未来电信”合作项目（PE00000001——“RESTART”计划）的支持。

设计无线边缘上的训练/推理任务，既采用协作方式，也采用分布式方式。这促成了所谓的边缘智能（EI）范式的定义[2]。在这一范式下，为实现预设的时延、准确率和能耗目标性能而进行的系统资源分配，已在 [3]–[6]中得到研究。具体而言，EI 使连接到移动网络的用户设备（UE）能够根据需要将其学习任务卸载到边缘服务器（ES），这些服务器部署在边缘侧，靠近无线接入点（RAP）。这使得系统资源（如传输速率、带宽和 CPU 时钟频率）能够依据特定优化策略得到高效管理，而这些策略主要聚焦于能耗、总体时延和学习准确率之间的权衡 [6]。

显然，从资源优化的角度来看，将完成学习任务所严格必需的最少信息量卸载到边缘服务器，同时满足性能要求并达到期望的准确率，将大有裨益。这一合理考量，加上未来 6G 网络 [7], 中预计将出现的流量大幅增长，促使人们寻求一种有别于经典香农设计的新型通信范式。在这一背景下，目标导向通信（GOC） [8] 是一个很有前景的选择。具体而言，如果通信的目的是对用户设备收集的数据执行推理任务，而非要求接收端准确重现所有传输比特，那么目标导向通信的目标就是只传输在边缘服务器执行推理任务时最相关的信息，同时保证规定的决策准确率和系统性能。这样既能帮助用户设备节省传输资源、避免不必要的数据速率增长，又能满足服务时延和能耗等应用约束。

相关工作。在 [6], [9], 中提出了开创性的 EI 框架，并采用无线卸载策略，通过动态分配用户设备向 ES 传输数据所使用的（量化）比特数来节省传输资源。在 [10] 和 [11], 中也采用了这种压缩策略，分别针对边缘分类和集成学习，并提供可靠性保证。一种更有理论依据、与学习任务更匹配的数据缩减策略，基于

信息瓶颈（IB） [12] [13], 已在 [14]中提出。然而，只有当整体统计数据服从联合高斯分布 [15] [14], 时，IB 原理才能为编码器提供闭式解；如果统计数据是离散的，则可以通过迭代机制获得解。当感知数据与决策输出既不服从联合高斯分布，也不是具有可控基数的离散数据时，推导 IB 解并不容易，因此，可以采用所谓的变分信息瓶颈（VIB）重新表述源编码问题，这一方法近期已在 [16] 和 [17], 中得到探讨，其中提出了一种协作式（多设备）推理框架。

联合信源信道编码（JSCC）为进一步偏离传统通信设计提供了一种可能性，随着深度神经网络（DNN）的广泛应用，这种方法日益受到关注。最近，已有多项研究提出用针对特定任务进行适当训练的 DNN，取代传统的信源编码器与信道编码器级联结构。例如， [18] 提出了一种基于 DNN 的 JSCC 方案，以在有限码块长度条件下提升图像检索应用的性能。此外，如果通信任务是图像识别，那么直接针对学习任务设计 JSCC 架构，而不是先重建图像再执行识别任务，是合理的； [19]提出了这种方法。 [20] 的作者提出了一种图像检索方案，将提取出的向量特征直接映射到信道输入符号，而不采用任何信道编码技术，并由服务器直接从含噪信道输出中检索最相关的图像。 [21],在此基础上扩展了该方法，在将提取出的特征映射到信道符号之前先对其进行量化。 [22], 将 JSCC 与在频率选择性信道上运行的 OFDM 系统相结合，而 [23] 则考虑将 JSCC 与非线性变换编码（NTC） [24]相结合。

就目标导向通信（也称任务导向通信）而言，近期已有多项研究证明了这一主题日益凸显的重要性。例如，在 [25] 和 [26] 中，研究人员利用目标导向压缩（ GOC ）定义了听者与说话者之间的<样式 id='3'>通用语言，并采用强化学习（RL）和课程学习（CL）；此外，文献 [27] 提出了一种基于 Transformer 的方法，用于辅助图像和文本传输。文献 [28],描述并评估了一种面向文本传输的噪声感知联合源信道编码（JSCC）方案，而文献[29]则利用混合自动重传请求（HARQ）方案来提高句子语义传输的可靠性。关于无人机（UAV）应用中的图像分类，以及目标导向压缩（GOC）辅助的视觉问答（VQA）任务，分别可参见文献 [30] 和[31], 。此外，文献 [19] 和 [20] 通过展示目标导向压缩方案在图像分类以及人和车辆重识别任务中带来的精度提升，论证了将 GOC 方案用于计算机视觉应用的必要性。最后，目标导向通信对语音识别任务的影响也已在文献 [32]中得到分析。

然而，上述引用的工作均未考虑多用户目标导向通信中数据缩减策略的动态优化，以及与全局网络资源管理的联合优化，同时还需满足既定性能要求。保证，正如我们在本文中所做的那样。沿着这一思路，在 [33]中，我们针对边缘辅助图像分类任务提出了最小能耗和最高准确率的资源分配策略，适用于单用户/单服务器场景；而在 [34] 中，我们报告了扩展到多用户场景的一些初步结果，并将在下文进一步发展和深入研究。

我们的贡献。 本文的主要贡献涉及系统架构、优化策略和仿真结果。具体总结如下：

1）系统架构： 在 [34], 中提出的初步策略基础上，我们考虑一种多用户目标导向通信场景，其中多个用户设备可以决定是否将其学习任务卸载到边缘服务器（ES）。每个用户使用一组源编码器，每个编码器对应一个特定的压缩比，并根据在线系统状态动态压缩要发送至 ES 的数据单元（DU）。具体而言，与 [33], 中一样，我们采用卷积编码器（CE），即卷积自编码器（CAE）的编码器，并通过一种新的训练函数提升其性能。ES 使用一组卷积分类器（CC），即 CNN，执行多个彼此独立于用户的推理任务，每个分类器都与用户设备端使用的 CE 相匹配。整个 CE-CC 结构有助于在用户设备与 *ES* 之间拆分分类任务。

2）优化策略： 我们对推理任务进行动态拆分，在每个时隙中，根据信道状态以及实时准确率和性能，从可用的（预训练）CE-CC中选出最合适的一对。具体而言，我们采用 Lyapunov 优化，实现一种面向多用户的动态目标导向源压缩架构，选择 CE-CC 对并分配计算和通信资源，在能耗（包括用户设备和边缘服务器的能耗）、时延与分类准确率之间进行权衡。下文中，我们在 [34], 中展示的初步结果和优化策略基础上，进一步考虑一种多用户最高准确率策略，同时保证（最大）时延上限和能耗（ MADE）。此外，我们允许每个用户设备自行决定是在本地执行推理任务，还是将其卸载到边缘服务器，因为有些应用可能具备足以在本地运行该应用的用户设备硬件，或者从整体资源管理的角度来看，本地执行可能更为合适。

3）仿真场景：我们研究了 [34], 中未分析的场景，其中每个用户设备都有不同的服务需求和约束。丰富多样的可能场景、优化策略和仿真结果，显著拓展了 [34], 中的研究成果，凸显了所提出的整体资源管理方法的有效性和灵活性。

大纲。本文组织结构如下。第 II 节阐述目标导向通信系统，以及用于分类的 CE 和 CC 联合训练过程。第 III 节介绍用于制定资源优化策略的整体系统模型，并在第 IV 节利用随机李雅普诺夫优化求解这些策略。第 V 节讨论我们的

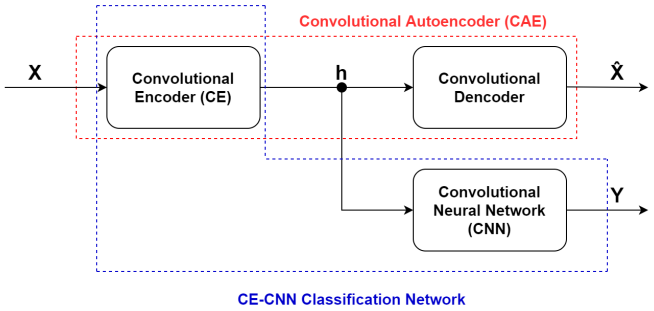

图 1：训练方案：CE h 的输出同时馈入 ES 分类 CNN 和一个 CD。

实验结果，最后在第 VI 节中得出一些结论，并强调未来的研究方向。

II. 分类网络与训练

本节介绍一种架构，旨在高效利用传输能耗和带宽。具体而言，在将用户设备的数据单元（DUs）（即学习任务的输入）传输至 ES 之前，我们会对其进行压缩。ES 必须在不牺牲预设目标准确率的前提下完成学习任务。正如 [33],中更详细的阐述，信息瓶颈（IB） [12] 是一种很有前景的理论框架，能够从目标导向的角度对数据源进行有意义的压缩。然而，IB 只有在相关统计量服从离散分布或高斯分布时才存在闭式解 [35] [14]。因此，鉴于我们关注的多分类图像分类任务并不满足高斯假设，且互信息难以进行有意义的定义 [36], ，我们在 [33]中提出了 IB 的一种启发式近似方法，该方法很好地契合了我们的目标导向策略。具体而言，我们的方法是在用户设备端部署可调的数据压缩机制，以服务于 ES 上相关的推理任务。在不限制整体 GOC 架构及其资源管理通用性的前提下，我们采用多个 CE，根据逐层最大池化策略在用户设备端压缩图像。CE 与 ES 端的 CC 配合完成最终决策，如图 1 所示，其中展示了单个用户设备的情形。

如 [33],中所述，交叉熵（CE）可实现为：

- **Short-CE：** 它通过一个卷积层后接一个最大池化层，将图像调整至所需分辨率。
- **Deep-CE：** 它通过多个卷积层对图像进行下采样，每个卷积层后都接一个最大池化层，将（伪）图像的尺寸减半。

请注意，我们的目标是对图像进行分类，而非重建。因此，对于图 1 所示的 CE-CC 压缩与分类网络，我们必须采用不同于经典 CAE 所用的学习代价函数。具体而言，我们采用以下目标函数

$$\underset{\theta,\phi}{\text{minimize}} \quad \frac{1}{N_t}\sum_{n=1}^{N_t} L_{ce}(Y_n,\hat{Y}_n,\phi,\theta)+\lambda L_{mse}(X_n,\hat{X}_n,\theta), \tag{1}$$

其中， $L_{ce}(Y_n, \hat{Y}_n, \phi, \theta)$ 是交叉熵损失，用于控制 ES 分类任务的性能；而 $L_{mse}(X_n, X_n, \theta)$ 是均方误差，用于衡量输入与完整 CAE 的重建版本 X 之间的差异。请注意，式 (1) 中的交叉熵损失是互信息 $I(h;Y)$ [37] 的代理。因此，在固定 CE 架构（压缩尺寸）的情况下，最小化交叉熵即可最大化 $I(h;Y)$，这构成了所提方法与 IB 原理之间的联系。然而，与我们在 [33], 中所做的不同，式 (1) 还考虑了卷积解码器（CD）的输出 MSE，即 CAE 中通常用于图像重建的部分。实际上，式 (1) 中加入这一（正则化）MSE 惩罚项，有利于提取有意义的特征 [38],，从而在参数 λ取值适当时提升整体学习任务的性能。无论如何，请注意，CD 仅在 CE-CC 训练期间纳入考虑，而不用于分类，这一点如图 1 所示。每一对 CE-CC 都需要经过适当训练，也可以由第三方离线训练。因此，尽管研究如何利用本文用于分类的同一无线边缘计算架构来训练分类网络很有意义，但这并非本文的研究内容，将留待未来研究。

JPEG 压缩。 为了获得良好的分类性能，交叉熵（CE）采用下采样原理压缩图像，这是由各层的最大池化策略实现的。然而，这种设计并未考虑用户设备（UE）与边缘服务器（ES）之间的无线通信。因此，尽管 CE 输出的潜在表示 h 的大小（见图 1）可能适合目标分类准确率，但相对于压缩数据单元的文件大小而言，仍可能不是最优的，从而导致传输能耗和时间成本高昂。这个问题说明，在将 h传输至 ES 之前，有必要进一步进行打包（压缩），并在 CC 输入端将其解包还原为 h 。鉴于分类任务的性质以及 CE 提取的伪图像 h 的结构，我们在用户设备端采用 JPEG 编解码器进行进一步压缩。事实证明，该编解码器能够有效减小数据单元的文件大小，同时从用户设备角度看，只需付出合理的额外计算开销。选择 JPEG 是因为它是一种广泛使用的打包系统，拥有大量高效的实现方案。此外，尽管 JPEG 是有损压缩，但已有研究证明，JPEG 编解码器不会显著影响 CNN 的分类性能 [39]。

III. 系统模型

所考虑的目标导向场景包含多个计算能力和能耗有限的设备（用户设备），这些设备通过一个接入点（AP）连接到具有更多计算资源的边缘服务器（ES）；具体示意如图2所示。为了执行通用学习任务，对于连接到网络的每个用户设备，系统需要处理三个主要阶段：i）用户设备缓存数据单元（DUs），即待分类的图像；ii）根据受系统状态影响的具体卸载决策，这些数据单元将被安排进行压缩和传输，由

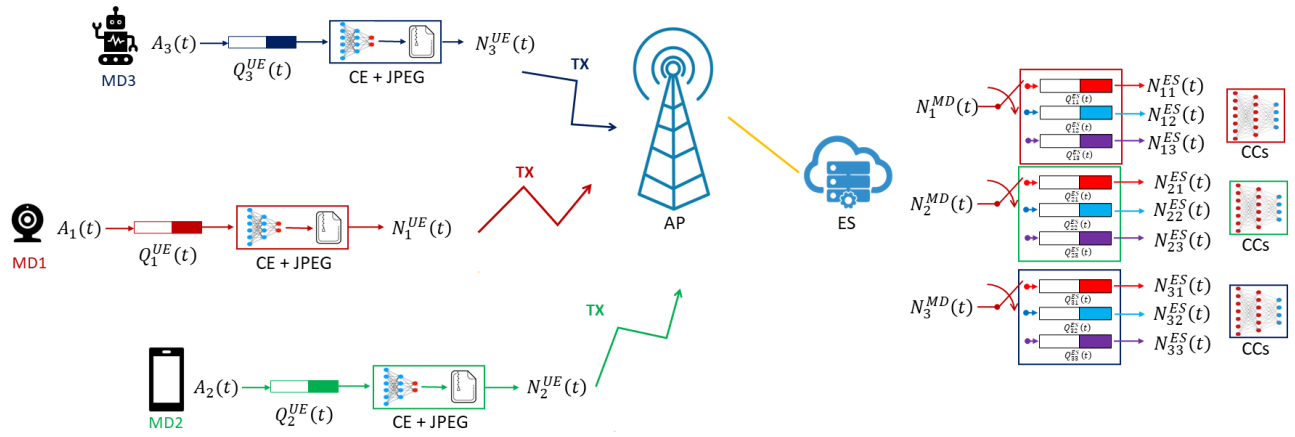

图2：场景：每个用户 U　E 动态地采用其自身的一组相互耦合的 CE　　在 ES 上设置一组合适的 CC。

第二节提出的目标导向压缩策略，或者在本地进行处理；iii) 推理任务根据卸载决策在用户设备端或边缘服务器端执行。

系统以时隙化方式演进，每个时隙的固定时长为 τ。因此，我们处理离散时间函数 $f(t)$，其中 $t \in \mathbb{N}$ 是第 t 个时隙 $[t\tau,(t+1)\tau[$ 的索引。GOC 的资源优化策略旨在满足特定的端到端（最大）时延要求，同时优化系统能耗或学习准确率。为此，所提出的策略必须管理多种资源。具体而言，第 k 个 UE 必须分配其面向 ES 的传输速率$R_k(t)$、用于以特定压缩因子$\rho_k(t)$执行数据压缩的时钟频率 $f_k^d(t)$，以及卸载决策 $d_k(t)$。对于 ES，主要优化变量是时钟频率 $f_c(t)$，它必须合理分配给不同用户的学习任务。这些量构成了我们将为所提出的资源管理策略定义的目标函数中的优化变量。现在我们可以介绍用于时延、能耗和分类准确率的模型。

A. 时延模型

根据李雅普诺夫优化框架 [40] 的规定，系统随时间的演化完全由一个排队系统描述。具体而言，对于网络中涉及的每个用户，我们定义两类物理队列：

- 每个 UE 都有一个计算/通信队列，用于收集每台设备生成的 DU，即图像；这些图像等待被压缩并传输至 ES 进行分类。- ES 侧另设一个计算队列，用于处理 UEs 可能动态采用的任意压缩等级（例如 CE）：因此，对于网络中的每个 UE，我们都有数量不同的 ES 队列，具体取决于可用的 CE 压缩等级。之所以采用这一设计，是为了使边缘服务器优化问题在计算上可承受，后文将对此作进一步说明。

我们用 K 表示连接到网络的用户设备总数。二元变量 $d_k(t) \in \{0,1\}$ 表示是否卸载第 k 个设备在第 t 个时隙的学习任务。当任一用户设备需要卸载其学习任务（即 $d_k(t) = 1$）时，我们作出以下假设，以便实际管理优化问题（更多详情请参见 [33]）。

假设*1*：每个用户设备（UE）队列中的数据单元（DUs）必须在同一个时隙内完成压缩和传输。事实上，在给定时隙内，无法对计划在后续某个时隙中传输的数据单元进行最优压缩，因为系统在后续时隙中可能会面临不同的信道条件，或者边缘服务器（ES）/用户设备（UE）的队列长度可能发生变化，等等。因此，压缩和传输操作必须在同一个时隙内按顺序执行。

假设*2*：我们假设，当用户设备正在传输一些数据单元时，也可以同时压缩其他数据单元。

在第 t个时隙内可能传输的（压缩后）数据单元数量表示为

$$N_k^{tx}(t) = \left\lfloor \frac{\tau R_k(t)}{M(\rho_k(t))N(\rho_k(t))} \right\rfloor, \tag{2}$$

其中，$R_k(t)$ 和 $\rho_k(t)$ 分别为时刻 t 为第 k 个 UE 选择的传输速率和压缩因子[1]，；$M(\rho_k(t))$ 是采用某一压缩因子 $\rho_k(t)$ 时数据单元的大小，而 $N(\rho_k(t))$ 是对（压缩后的）伪图像 h 中一个像素进行编码所需的平均比特数。为简化记号，我们还定义$W(\rho_k(t)) = M(\rho_k(t))N(\rho_k(t))$，表示存储具有给定 $\rho_k(t)$的图像所需的平均比特数。另一方面，可行的数据单元数量 $N_k^c(t)$

[1]注意，我们用 ρ 表示图像在每个维度上的压缩因子。实际压缩比随 ρ^2变化。

由第 k 个设备在第 t 个时隙进行压缩的表示为

$$N_k^c(t) = \lfloor \tau f_k^d(t) J_d(\rho_k(t)) \rfloor, \tag{3}$$

其中，$J_d(\rho_k(t))$ 表示在一个时钟周期 C 内压缩的数据单元数量（该数量取决于所选的压缩因子 $\rho_k(t)$），而 $f_k^d(t)$ 表示在同一时隙内为第 k 个用户设备选定的设备时钟频率。根据假设1，在一个时隙内压缩的所有数据单元都必须在同一时隙内传输，并且所有待传输的数据单元都必须先经过压缩。因此，我们需要采用传输速率 $R_k(t) \leq W(\rho(t)) f_k^d(t) J_d(\rho(t))$，从而得到 $N_k^{tx}(t) \leq N_k^c(t)$。考虑到在传输开始之前，我们需要等待一段等于 $1/(f_k^d(t) J_k^d(t))$ 的时间来压缩第一个数据单元，因此，第 k 个设备在第 t 个时隙内实际能够卸载的数据单元数量表示为

$$N_k^{off}(t) = \left\lfloor \frac{\tau - 1/(f_k^d(t) J_k^d(t))}{W(\rho_k(t))/R_k(t)} \right\rfloor. \tag{4}$$

将不等式 $N_k^{tx}(t) \leq N_k^c(t)$ 代入（4），可得以下（整数）不等式

$$\left\lfloor \frac{\tau R_k(t)}{W(\rho_k(t))} \right\rfloor - 1 \leq N_k^{off}(t) \leq \left\lfloor \frac{\tau R_k(t)}{W(\rho_k(t))} \right\rfloor, \tag{5}$$

这将在接下来的推导中发挥作用。

最后，与（3）类似，当学习任务在本地执行时，第 k 个用户设备处理的数据单元总数表示为

$$N_k^L(t) = \lfloor \tau f_k^d(t) J_k^L(\rho_k(t)) \rfloor, \tag{6}$$

其中，$J_k^L(\rho_k(t))$ 表示用户设备硬件在一个时钟周期内可压缩的数据单元数量，压缩因子为 $\rho_k(t)$，并随后对其进行分类。结合（4）和（6），用户设备在单个时隙内可处理的数据单元数量表示为

$$N_k^{UE}(t) = d_k(t) \cdot N_k^{off}(t) + (1 - d_k(t)) \cdot N_k^L(t). \tag{7}$$

用户设备队列 $Q_k^{UE}(t)$ 由新数据单元的到达进行填充，并通过将数据单元传输至边缘服务器或在用户设备本地对其进行分类而排空。因此，其演变过程如下

$$Q_k^{UE}(t+1) = \max(0, Q_k^{UE}(t) - N_k^{UE}(t)) + A_k(t), \tag{8}$$

其中，$A_k(t)$ 表示数据单元到达过程，其统计特性通常未知。

在边缘服务器处，我们为每个用户设备采用 L_k 个不同的队列，其演化过程描述为

$$\begin{aligned} Q_{ki}^{ES}(t+1) = \max(0, Q_{ki}^{ES}(t) - N_{ki}^{ES}(t)) \\ + d_k(t) \cdot \min(N_k^{UE}(t), Q_k^{UE}(t)) \cdot \mathbb{1}_i\{\rho_k(t)\}, \end{aligned} \tag{9}$$

即，为集合 $\mathcal{S}_k = \{s_{ki}\}_{i=1,\ldots,L_k}$ 中的 L_k 各设一个队列，该集合表示第 k 个 UE 可采用的压缩因子集合。这些队列存储为第 k 个设备预留的 ES 计算负载，以 DU 数量表示。术语 $\mathbb{1}_i\{\rho_k(t)\}$ 中 (9) 是指示函数 $\mathbb{1}\{\rho_k(t) = s_{ki}\}$ 的简写，它表示仅当 UE 选择第 i 个压缩因子时，新的 DU 才会到达 ES 队列。术语 $N_{ki}^{ES}(t)$ 在 (9) 中表示 ES 在第 t 个时隙内处理的 DU 数量，其表达式为

$$N_{ki}^{ES}(t) = \lfloor \tau f_{ki}^s(t) J_{ki}^s(t) \rfloor, \tag{10}$$

其中，$f_{ki}^s(t)$ 表示在第 t 个时隙内，分配给第 k 个用户设备的第 i 个队列（压缩因子）的边缘服务器时钟频率。[2] 式（10）中的量 $\frac{1}{J_{ki}^s(t)}$ 是一个转换因子，用于将边缘服务器接收的数据单元数量转换为处理这些数据单元（例如进行分类）所需的等效时钟周期数。

为设定时延约束，我们需要定义一个总体队列，以针对每台设备综合考虑用户设备侧和边缘服务器侧的总体计算负载。由于我们的目标是满足平均时延约束（下文将对此进行详细说明），并且考虑到边缘服务器能够并行处理多个数据单元，因此，根据式（8）和式（9），考虑并行队列的平均长度是合理的，其表达式为

$$Q_k^{tot}(t) = Q_k^{UE}(t) + \sum_i^{L_k} p_{ki} Q_{ki}^{ES}(t), \tag{11}$$

其中，p_{ki} 是在 $\mathcal{S}_k$ 中采用第 i 个压缩因子的概率，可通过在线样本均值 [3] 进行估计。假设数据到达率为 $A_k = \overline{\quad}$，并利用利特尔法则（ ）可对长期平均时延进行建模，其表达式为

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left\{ \frac{Q_k^{tot}(t)}{\overline{A_k}} \right\}. \tag{12}$$

对于时延约束 D_k^{avg}，我们得到队列长度约束 $Q_k^{avg} = D_k^{avg} A_k$，因此可以将时延约束等价地形式化为队列约束：

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\{Q_k^{tot}(t)\} \leq Q_k^{avg}. \tag{13}$$

B. 能耗模型

我们的系统能耗模型包含三个主要组成部分：

- 用户设备的传输能耗，在作出卸载决策时，用户设备需要将数据单元传输至边缘服务器。

[2] 为每个压缩因子设置不同的队列是一种设计选择，有助于建立 N_{ki}^{ES} 与 f_{ki}^s 之间的数学关系，且该关系比在 [33], 中使用单个队列时更简单。这样，即使在多用户场景下，边缘服务器优化问题的求解也变得可行，稍后我们将对此作进一步说明。

[3] 这些 $p_{k,\ i}$ 实际上会随系统状态而变化，同时也受到我们最终采用的瞬时和自适应资源管理策略的影响。这里的假设是，利用对估计的 $p_{k,\ i}$ 的了解的随机资源管理算法，将收敛到一个稳态，在该稳态下，$p_{k,\ i}$ 的运行样本均值估计也将收敛。这一事实已通过大量仿真结果得到验证。

- 用户设备的计算能耗，用于压缩/编码待传输的数据单元，或在本地执行学习任务。- 边缘服务器的计算能耗，用于对决定卸载学习任务的用户设备所传输的数据单元进行分类。

为简便起见，假设采用容量达到型传输系统，在平坦衰落无线信道中，第 k 个用户设备所请求的传输功率 $p_k^{tx}(t)$ 可根据香农容量 [42]推得

$$R_k(t) = B_k \log_2 \left(1 + \frac{p_k^{tx}(t)|h_k(t)|^2}{N_0 B_k}\right), \tag{14}$$

其中， $|h_k(t)|$ 表示信道增益， N_0 表示接收端的噪声功率谱密度， B_k 表示带宽。因此，通过对式（14）求逆可得，第 k 个用户设备在第 t 个时隙内产生的传输能耗取决于速率 $R_k(t)$，具体为

$$E_{tx}^k(t) = \tau p_k^{tx}(t) = \frac{\tau B_k N_0}{|h_k(t)|^2}\left(e^{\frac{R_k(t)ln(2)}{B_k}} - 1\right). \tag{15}$$

从计算角度来看，边缘服务器和用户设备的模型是等价的。具体而言，为估计能耗，我们采用 [43],中的模型，该模型假设边缘服务器和用户设备的时钟频率$f_s(t)$和 $f_k^d(t)$之间呈立方关系，如下式所示

$$E_k^d(t) = \tau \kappa_k^d f_k^d(t)^3 \quad \text{and} \quad E_s(t) = \tau \kappa_s f_s(t)^3. \tag{16}$$

常数 κ_s 和 κ_k^d 分别表示边缘服务器和第 k 个用户设备处理器的有效开关电容 [43] 。因此，我们使用以下加权性能指标来量化第 t 个时隙内的系统能耗：

$$E_k^{tot}(t) = (1-\gamma)E_s(t) + \gamma \sum_{k=1}^{K} \delta_k (E_k^c(t) + E_k^{tx}(t)), \tag{17}$$

其中，参数 γ 用于对比用户设备与边缘服务器的能耗进行加权，从而可以调整优化策略，使其以用户为中心（ $\gamma \to 1$）或以服务器为中心（$\gamma \to 0$）。此外，权重 $\{\delta_k\}_{k=1}^K$ （其中 $\sum_{k=1}^K \delta_k = 1$）可用于为不同用户的能耗赋予不同的重要性，根据运营商、用户和服务提供商的需求，为优化提供额外的灵活性。

C. 准确率模型

对于每个用户设备的学习任务准确率，我们采用基于模型的管理策略。这意味着，第 k 个任务的准确率可以在优化问题中表示为压缩程度的函数 $G_k(\rho_k(t))$。实际中可以通过使用查找表（LUT）（见第 V 节）来实现，其中每个表项都与特定的压缩因子 $\rho_k \in \mathcal{S}_k$相关。[4] 该 LUT 存储第 k 个学习任务的（平均）分类准确率，并将其与第 k 个用户设备可用的每条 CE-CC 分类链相关联。在每种 CE-CC 结构经过适当训练后，可以按照上一节所述，利用有意义的测试集离线估算准确率 LUT 中存储的值。因此，我们可以利用 LUT $G(\rho_k(t))$，对每个学习任务施加平均准确率约束，如下所示：

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\{G_k(\rho_k(t))\} \geq G_k^{avg}. \tag{18}$$

IV. 面向目标导向通信的动态资源优化

多用户目标导向通信

基于上一节提出的时延、准确率和能耗模型，我们设计了两种资源优化策略：具有（最大）时延和准确率约束的多用户最小能耗策略（mu-MEDA），以及具有（最大）时延和能耗约束的多用户最大准确率策略（mu-MADE）。下文将介绍这两种策略的问题表述和算法求解方案。

A. mu-MEDA：具有时延和精度约束的多用户最小能耗

从系统能耗最小化的角度出发，长期优化问题可表述如下：

$$\begin{aligned}
\min_{\Phi(t)} \quad & \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\{E_{tot}(t)\} \\
\text{s.t.} \quad & (a)\ \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\{Q_k^{tot}(t)\} \leq Q_k^{avg}, \forall k \\
& (b)\ \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\{G(\rho_k(t)\} \geq G_k^{avg}, \forall k \\
& (c)\ 0 \leq R_k(t) \leq R_{k,max}, \quad \forall k,t \\
& (d)\ \rho_k(t) \in \mathcal{S}_k, \quad f_s(t) \in \mathcal{F}_s, \quad f_k^d(t) \in \mathcal{F}_{d,k} \quad \forall k,t \\
& (e)\ \sum_{k=1}^{K}\sum_{i=1}^{L_k} f_{ki}^s(t) \leq f_s(t), \quad (f)\ f_{ki}^s(t) \geq 0 \\
& (g)\ d_k(t) \in \{0,1\} \quad \forall k,i,t
\end{aligned} \tag{19}$$

其中， $\Phi(t) = [\{R_k(t)\ f_{ki}^s(t)\ f_k^d(t)\ \rho_k(t)\ d_k(t)\}_{k=1}^K, f_s(t)]$包含所有优化变量。约束条件(19)的含义如下：(*a*) 第 k个用户设备的平均队列长度必须低于 Q_k^{avg} ，即我们设定的最大平均服务时延为 $D_{avg}^k = Q_{avg}^k / A_k$(参见(13))；(*b*) 第~~k~~个用户设备的平均分类准确率】【。

[4]我们通过查找表（LUT）对压缩因子与准确率之间的关系进行建模，而非采用正式的解析表达式，因为在实践中几乎不可能找到该函数的闭式表达式。事实上，尽管 [44], [45],这些方法中有一些显著的实例试图从理论上形式化 DNN 的性能，但它们基于互信息，而在大多数实际情况下，互信息难以推导出闭式表达式。

UE 必须大于 G_{avg}^{k}；(c) 第 k个 UE 的传输速率 $R_k(t)$必须小于数值$R_{k,\ max}(t)$，该数值是根据式（14）并考虑最大可用传输功率$p_{k,\ max}^{tx}$推得的第 k个设备的最大可能速率；(d)规定了离散集合 $\mathcal{F}_c$、 $\mathcal{F}_{d,\ k}$ 和$\mathcal{S}_k$ ，它们分别表示服务器频率集合、第k个 UE 的频率集合以及可能的压缩因子集合；约束(e) – (f)指出，边缘服务器为分配给每个用户的所有队列分配的时钟频率之和 $f_{ki}^{s}(t)$，必须低于第 t 个时隙所选的 ES 总时钟频率，并且每个时钟频率显然都必须大于 0；最后，(g)表示对每个 UE 的机会式卸载决策变量集合施加的二元约束。由于缺乏无线信道和数据到达过程的统计信息（而这些信息是计算式（19）中期望值所必需的），问题（19）较为复杂。为解决这一问题，我们采用 Lyapunov 随机优化方法 [40]，，将长期问题（19）转化为一系列瞬时优化问题，从而可以按时隙求解。根据这一优化框架 [40],，我们首先为每个长期约束(a)和(b)分别关联一个虚拟队列。这些虚拟队列按照以下方式演化

$$\begin{aligned} Z_k(t+1) &= \max(0, Z_k(t) + \mu_k(Q_k^{tot}(t+1) - Q_k^{avg})) \\ Y_k(t+1) &= \max(0, Y_k(t) + \nu_k(G_k^{avg} - G_k(t))), \end{aligned} \tag{20}$$

其中， μ_k 和 ν_k 是控制算法收敛速度的步长。由此可证明，满足长期约束（a） –（b）等价于保证式（20）[40] 中虚拟队列的均值速率稳定性。为此，我们将李雅普诺夫函数定义为 $L(t)$，即所有（虚拟和物理）队列的平方和。

$$L(t) = \sum_{k=1}^{K} Z_k(t)^2 + \sum_{k=1}^{K} Y_k(t)^2. \tag{21}$$

定义 $\Theta(t) = \left[\{Z_k(t)\}_{k=1}^{K}, \{Y_k(t)\}_{k=1}^{K}\right]$，我们得到相应的条件*Lyapunov* 漂移

$$\Delta(\Theta(t)) = \mathbb{E}\{L(t+1) - L(t) | \Theta(t)\}, \tag{22}$$

其最小化对应于虚拟队列的稳定化，但并未考虑目标函数（即系统能耗）。因此，为了权衡系统稳定性与能耗，在 Lyapunov 漂移中增加一个依赖于系统能耗的项，从而得到所谓的 *Lyapunov* 漂移加惩罚项函数

$$\Delta_p(\Theta(t)) = \Delta(\Theta(t)) + V\mathbb{E}\{E_{tot}(t)\}. \tag{23}$$

通过增大参数 V 的取值，我们会更加重视目标函数，而非队列稳定性，从而推动解趋于最优，同时仍保证系统稳定，即满足长期约束。具体而言， [40] 证明，随着参数 V 增大，(19) 的最优解渐近可达。根据随机优化论证， [40], 我们继续最小化 (23) 中 Lyapunov 漂移加惩罚函数的一个上界（推导见附录），最终得到 (24) 中的瞬时优化问题。由于优化变量仅影响项 N_k^{UE}、 N_{ki}^{ES} 和 G_k，我们忽略所有与这些变量无关的项。还需注意，为简化记号，以下省略时间索引 t 。

$$\begin{aligned} \min_{\Phi} \quad & VE_{tot} + \sum_{k=1}^{K}\Bigg[L_k N_k^{UE}\mu_k^2\left(\sum_{i=1}^{L_k}\mathbb{1}_i\{\rho_k\}p_{ki}Q_{ki}^{ES} - Q_k^{UE}\right) \\ & - L_k\mu_k^2\sum_{i=1}^{L_k}p_{ki}Q_{ki}^{ES}N_{ki}^{ES} + \mu_k Z_k(\max(0, Q_k^{UE} - N_k^{UE}) \\ & + \sum_{i=1}^{L_k}\max(0, p_{ki}Q_{ki}^{ES} - N_{ki}^{ES})) - \nu_k Y_k G_k(\rho_k)\Bigg] \\ \text{s.t.} \quad & 0 \le R_k \le R_{k,max}, \quad \rho_k \in \mathcal{S}_k, \quad f_s \in \mathcal{F}_s, \quad f_k^d \in \mathcal{F}_{d,k} \\ & \sum_{k=1}^{K} f_k^s \le f_s, \quad f_k^s \ge 0, \quad \forall k, t. \end{aligned} \tag{24}$$

由于问题（24）的代价函数中的用户设备能耗项仅取决于用户设备的优化变量 $\{\Phi_{d,k}\}_{k=1}^{K} = \{[R_k, f_k^d, \rho_k, d_k]\}_{k=1}^{K}$（且各用户设备之间相互独立），因此我们可以在每个用户设备处分别对代价函数的这一部分进行优化。值得注意的是，我们在边缘服务器处为每个用户设备的卸载任务分别分配独立计算队列的设计选择，使我们能够完全解耦该优化问题，并分别处理用户设备和边缘服务器的资源优化。此外，正如脚注2中所指出的，针对每个压缩因子ρ_{ki}使用多个队列，借助（11），使得问题（10）相对于 f_{ki}呈线性，直至 $\lfloor . \rfloor$ 算子。因此，问题（24）具有可分离性，可以针对每个压缩因子分别求解，具体如下所述。

1) 用户设备子问题：对于第 k 个设备，在每个时隙t，我们必须求解以下优化问题。

$$\begin{aligned} \min_{\Phi_{d,k}} \quad & L_k N_k^{UE}\mu_k^2\left(\sum_{i=1}^{L_k}\mathbb{1}_i\{\rho_k\}p_{ki}Q_{ki}^{ES} - Q_k^{UE}\right) \\ & + \mu_k Z_k \max(0, Q_k^{UE} - N_k^{UE}) - \nu_k Y_k G_k(\rho_k) \\ & + V\gamma\delta_k(E_k^{tx} + E_k^c) \qquad (25) \\ \text{s.t.} \quad & 0 \le R_k \le R_{k,max}, \quad \rho_k \in \mathcal{S}_k, \quad f_k^d(t) \in \mathcal{F}_{d,k}, \qquad (26) \\ & d_k \in \{0, 1\}. \end{aligned}$$

根据卸载决策变量d_k 的取值，我们可以采用两种不同的策略来优化其他变量。如果 $d_k = 1$，我们必须分配传输速率 R_k 以将数据单元传输至边缘服务器，并分配用户设备时钟频率 f_k^d 和压缩因子 ρ_k 来执行压缩。否则，如果 $d_k = 0$ ，我们只需分配f_k^d 和 ρ_k ，以便在本地执行学习任务。我们指出，尽管这并非强制要求，但我们假设用户设备在本地也采用为 GOC 方案设计的相同（组）交叉熵-交叉熵分类链，从而公平地为用户设备提供与边缘服务器方案所能利用的相同分类准确率和能耗灵活性。

分类器在用户设备（UE）上的其他选择或固定结构，显然会影响最优资源管理所做出的卸载决策，进而影响能耗-时延-准确率之间的权衡。

关于该问题的求解，当我们通过对第 k个用户的传输速率添加以下约束来处理（4）中的 min(·) 时

$$\begin{aligned} &0 \le R_k(t) \le R_{k,max}^{+}(t), \\ &R_{k,max}^{+} = \min\left\{R_{k,max}, \frac{Q_k^{UE} W(\rho_k)}{\tau}\right\}. \end{aligned} \tag{27}$$

因此，根据假设 1 和 2，并考虑到我们无法压缩超过可传输数量的数据单元，我们选择的数据速率受以下两者中的较小值约束：最大可实现传输速率$R_{k,\ max}$ （将最大功率 p_k^{tx} 代入香农容量（ 14 ）计算得到），以及排空速率 $Q_k^{UE} W(\rho_k)/\tau$ ，后者能够清空传输队列（并去除 max(·)）。考虑到 $x-1 \le \lfloor x \rfloor \le x$，我们还可以去除（4）中的 $\lfloor\cdot\rfloor$。因此，利用指示函数的定义，对于任意固定的压缩因子 $\rho_{ki} \in \mathcal{S}_k$，最终得到以下优化问题

$$\begin{aligned} \min_{\Phi_{d,k}} \quad & -\frac{Q_{ki}^{TX}\tau R_k}{W(\rho_k)} + \frac{\tau V\gamma\delta_k B_k N_0}{h_k^2} e^{\frac{R_k ln(2)}{B_k}} + \tau V\gamma\delta_k\kappa(f_k^d)^3 \\ & - \nu_k Y_k G_k(\rho_k) \\ \text{s.t.} \quad & 0 \le R_k \le R_{k,max}^{+}, \quad f_k^d \in \mathcal{F}_{d,k}, \end{aligned}$$

其中 $Q_{ki}^{TX} = L_k\mu_k^2(Q_k^{UE} - p_{ki}Q_{ki}^{ES}) + \mu_k Z_k$。这是一个混合整数优化问题。然而，在实践中，集合$\mathcal{F}_{d,\ k}$ 和 S_k 的基数都很小，因此，如下文所述，可以通过穷举搜索快速找到解。事实上，对于任意固定的压缩因子 $\rho_k \in \mathcal{S}_k$ 和计算频率 $f_k^d \in \mathcal{F}_{d,\ k}$，关于数据速率$R_k$的优化问题是凸的，其最优值可通过对偶理论和拉格朗日方法以闭式求得

$$\begin{aligned} \mathcal{L} = & -\frac{\tau Q_{ki}^{TX} R_k}{M(\rho_k)N(\rho_k)} + \frac{\tau V\gamma\delta_k N_0 B_k}{h_k^2} e^{\frac{R_k ln(2)}{B_k}} + \tau V\gamma\delta_k\kappa(f_k^d)^3 \\ & - \nu_k Y_k G_k(\rho_k) - \alpha R_k + \beta(R_k - R_{k,max}^{+}), \end{aligned} \tag{28}$$

其中， α 和 β 是拉格朗日乘子。注意，如果 $Q_{ki}^{TX} \le 0$，第二项会随速率单调递增，因此当 R_k =0 时，拉格朗日函数取得最小值。否则，当 $Q_{ki}^{TX} > 0$ 时，我们可以通过施加以下 KKT 条件来求解优化问题： [46]

$$\begin{aligned} &(a)\ \frac{\partial\mathcal{L}}{\partial R_k} = -\frac{Q_{ki}^{TX}\tau}{W(\rho_k)} + \frac{\tau V\gamma\delta_k ln(2) N_0 B_k}{h_k^2} e^{\frac{R_k ln(2)}{B_k}} \\ &- \alpha + \beta = 0 \\ &(b)\ 0 \le R_k \le R_{k,max}^{+},\ (c)\ \alpha \ge 0,\ (d)\ \beta \ge 0 \\ &(e)\ \alpha R_k = 0,\ (f)\ \beta(R_k - R_{k,max}^{+}) = 0. \end{aligned} \tag{29}$$

求解 KKT 条件可得以下方程，用于计算最优速率

$$R_k^*(\rho_k, f_k^d) = \begin{cases} \left[\frac{B_k}{\ln(2)} \ln\left(\frac{Q_{ki}^{TX} h_k^2}{W(\rho_k) V\gamma\delta_k \ln(2) N_0}\right)\right]_0^{R_{max}^{+}} & Q_{ki}^{TX} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

由此可得，在压缩因子 ρ_k和第 k 个用户的时钟频率 f_k^d 固定时，最优速率的闭式表达式。因此，正如前文所述，得益于$\mathcal{F}_{d,\ k}$和 $\mathcal{S}_k$的基数有限，我们可以通过穷举搜索选出最佳时钟频率 f_k^{d*}和压缩因子 ρ_k^*。总之，对于可能的卸载情形（$d_k = 1$），我们针对每个可能的压缩因子 ρ_k 计算最优速率和时钟频率 f_k^d ，然后在每个时隙选择使能耗代价最低的三元组 $T_k^* = (R_k^*,\ f_k^{d*},\ \rho_k^*)$。否则，对于在 UE 端进行分类的情形（$d_k = 0$），向 ES 的传输速率为 $R_k = 0$，我们只需针对每个可能的压缩因子优化时钟频率，从而得到使 UE 能耗最小的最优数对$P_k^* = (f_k^{d*},\ \rho_k^*)$。UE 优化问题的整体最优解还包括是否卸载学习任务的决策，最终通过在数对 $(d_k = 1,\ T_k^*)$ 和 $(d_k = 0,\ P_k^*)$ 中进行选择，取使 UE 能耗代价函数值最小的一组。

2）边缘服务器子问题：从边缘服务器的角度来看，对于每个用户设备，我们都必须管理多个计算队列，每个计算队列对应于该用户设备所采用的特定压缩因子：下文中，我们用Q_{ki}^{ES}表示第 i个用户设备的第 k个边缘服务器计算队列。显然，将分配给第 k个用户的第 i个队列的计算频率份额 f_{ki}^s（占边缘服务器总计算频率 f_s的比例）限制为不超过在一个时隙内完全清空该队列所需的份额是合理的，具体如下

$$f_{ki}^s(t) \le \min\left(f_s(t), \frac{Q_{ki}^{ES}(t)}{\tau J_{ki}^s(t)}\right). \tag{31}$$

这样，我们就可以从式（24）的求和中移除max(0, $Q_{ki}^{ES} - N_{ki}^{ES}$)项，因此可以将边缘服务器的资源分配问题改写为

$$\begin{aligned} \min_{\Phi_s} \quad & -\sum_{k=1}^{K}\sum_{i=1}^{L_k} \tau Q_{ki}^{comp} J_{ki}^s f_{ki}^s + \tau V(1-\gamma)\kappa f_s^3 \\ \text{s.t.} \quad & 0 \le f_{ki}^s(t) \le \min\left(f_s, \frac{Q_{ki}^{ES}}{\tau J_{ki}^s}\right) \\ & \sum_{k=1}^{K}\sum_{i=1}^{L} f_{ki}^s \le f_s, \quad f_s \in \mathcal{F}_s, \end{aligned} \tag{32}$$

其中， $\Phi_s = [\{f_{ki}^s\}_{i=1,\ldots,L_k,k=1,\ldots,K}, f_s]$，且 $Q_{ki}^{comp} = L_k\mu_k^2 Q_{ki}^{ES} + \mu_k Z_k$。尽管该问题是一个混合整数优化问题，但对于任意固定的边缘服务器时钟频率 f_s，它都可归结为经典的（分数）背包问题 [47]。因此，可通过贪心算法获得最优解：按照队列的权重（$Q_{ki}^{comp} J_{ki}^s$）降序排列，然后将时钟频率分配给队列，分配值为 min $\left(\phi, \frac{Q_{ki}^{ES}}{\tau J_{ki}^s}\right)$，其中 ϕ

是边缘服务器时钟频率 $f_c(t)$ 的剩余部分。由于边缘服务器时钟频率集合 $\mathcal{F}_s$ 的基数有限，在这种情况下，我们同样可以针对所有服务器时钟频率 $f_s \in \mathcal{F}_s$ 穷举求解该问题，从而获得可能的解集 $\{(f_{ki}^s, f_s)\}_{f_s \in \mathcal{F}_s}$，然后选择与式 (32) 中边缘服务器成本最小值相对应的解。

B. mu-MADE：延迟和能耗约束下的多用户最高准确率

另一种以最高准确率为目标的资源分配方案可表述为

$$
\begin{aligned}
\min_{\Phi(t)} \quad & \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\{\sum_{k=1}^{K}-G_k(t)\right\} \\
\text{s.t.} \quad & (a)\ \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{Q_k^{tot}(t)\} \le Q_k^{avg} \\
& (b)\ \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{E_k^{d}(t)\} \le E_k^{d,avg} && \forall k \\
& (c)\ \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{E_s(t)\} \le E_s^{avg} \\
& (d)\ 0 \le R_k(t) \le R_{k,max} && \forall k,t \\
& (e)\ \rho_k(t)\in\mathcal{S}_k,\ f_s(t)\in\mathcal{F}_s,\ f_k^d(t)\in\mathcal{F}_{d,k} && \forall k,t \\
& (f)\ \sum_{k=1}^{K}\sum_{i=1}^{L_k} f_{ki}^s(t) \le f_s(t) && \forall k,t \\
& (g)\ f_{ki}^s(t) \ge 0 && \forall k,i,t \\
& (h)\ d_k(t)\in\{0,1\} && \forall k
\end{aligned}
\tag{33}
$$

其中，$\Phi(t) = [\{R_k(t)\ f_{ki}^s(t)\ f_k^d(t)\ \rho_k(t)\ d_k(t)\},\ f_s(t)]$，对于 $k = 1, \ldots\ldots, K$，以及 $i = 1, \ldots\ldots, L_k$ 包含所有优化变量。(19)中的约束含义如下：(a)第 k 个 UE 的平均队列长度必须低于 Q_k^{avg}，即我们将最大平均服务时延设为 $D_{avg}^k = Q_{avg}^k/A_k$（参见(13)）；($b$)第 k 个 UE 的平均能耗必须低于 $E_{d,\ avg}^k$；(c) ES 的平均能耗必须低于 E_{avg}^s；(d)–(h)与(19)中的(c)–(g)等价。

与 mu-MEDA 策略类似，为了管理长期能耗约束 (b) 和 (c)，除了在式 (20) 中定义的用于管理 (a) 的虚拟队列 $Z_k(t)$ 外，我们还需要定义以下虚拟队列

$$
\begin{aligned}
S_k(t+1) &= \max(0, S_k(t) + \lambda_k(E_k^d(t+1) - E_k^{d,avg})) \\
O(t+1) &= \max(0, O_k(t) + \eta(O(t+1) - E_s^{avg})),
\end{aligned}
\tag{34}
$$

其中，$\{\lambda_k\}_{k=1}^K$ 和 η 是用于控制算法收敛速度的步长。根据虚拟队列的定义，在这种情况下，李雅普诺夫函数变为

$$
L(t) = \sum_{k=1}^{K}[S_k(t)^2 + Z_k(t)^2] + O(t)^2 \tag{35}
$$

因此，给定 $\Theta(t) = [\{S_k(t)\ Z_k(t)\}_{k=1}^K,\ O(t)]$，我们推导出以下 Lyapunov 漂移加惩罚函数的表达式。

$$
\Delta_p(t) = \mathbb{E}\{L(t+1) - L(t)|\Theta(t)\} - V\mathbb{E}\left\{\sum_{k=1}^{K}G_k(t)\right\} \tag{36}
$$

如附录中所述，最终得到以下优化问题

$$
\begin{aligned}
\min_{\Phi} \sum_{k=1}^{K}\Bigg[& L_k N_k^{UE}\mu_k^2\left(\sum_{i=1}^{L_k}\mathbb{1}_i\{\rho_k\}p_{ki}Q_{ki}^{ES} - Q_k^{UE}\right) \\
& + \mu_k Z_k(\max(0, Q_k^{UE} - N_k^{UE}) + \sum_{i=1}^{L_k}\max(0, p_{ki}Q_{ki}^{ES} - N_{ki}^{ES})) \\
& + \lambda_k S_k E_k^d - L_k\mu_k^2\sum_{i=1}^{L_k}p_{ki}Q_{ki}^{ES}N_{ki}^{ES}\Bigg] + \eta O E_k^s - V\sum_{k=1}^{K}G_k(t) \\
\text{满足 } & 0 \le R_k \le R_{k,\ max}, \quad \rho_k \in \mathcal{S}_k, \quad f_s \in \mathcal{F}_s, \quad f_k^d \in \mathcal{F}_{d,\ k} \\
& \sum_{k=1}^{K} f_k^s(t) \le f_s(t),\ f_k^s \ge 0, \quad \forall k,t.
\end{aligned}
\tag{37}
$$

再次利用所提设计所实现的问题解耦，分别处理每个特定 UE 和每个特定压缩因子的队列，我们在此情形下也得到彼此独立的瞬时优化问题：每个 UE 对应一个，ES 对应一个。

1) 用户设备子问题：对于第 k 个 UE，我们得到以下优化问题的形式

$$
\begin{aligned}
\min_{\Phi_{d,k}} \quad & L_k N_k^{UE}\mu_k^2\left(\sum_{i=1}^{L_k}\{\mathbb{1}_i\{\rho_k\}p_{ki}Q_{ki}^{ES}Q_k^{UE}\right) \\
& + \mu_k Z_k\max(0, Q_k^{UE}N_k^{UE}) \\
& - \lambda_k S_k E_k^d - V G_k(\rho_k) \\
\text{满足 } & 0 \le R_k(t) \le R_{k,\ max} \\
& \rho_k(t)\in\mathcal{S}_k,\ f_k^d(t)\in\mathcal{F}_{d,k},\ d_k\in\{0,1\},\ \forall k,t
\end{aligned}
\tag{38}
$$

其中 $\Phi_{d,\ k} = [R_k, f_k^d, \rho_k, d_k]$，对于 $k = 1, \ldots\ldots, K$。该问题的求解策略与前一种能耗最小化情形非常相似：如果 UE 决定卸载其任务（$d_k = 1$），我们需要在压缩因子 ρ_k 和设备时钟频率 f_k^d 固定的情况下，为其分配最优传输速率 R_k。同样，在这种情况下，我们可以通过对偶理论求得闭式最优速率

$$
R_k^*(\rho_k, f_k^d) = \begin{cases} \left[\frac{B_k}{\ln(2)}\ln\left(\frac{Q_{ki}^{TX}h_k^2}{W(\rho_k)\lambda_k S_k\ln(2)N_0}\right)\right]_{,}^{R_{max}^+} & Q_{ki}^{TX} > 0 \\ 0 & \text{otherwise} \end{cases}
\tag{39}
$$

因此，对于一个可能的卸载决策（$d_k = 1$），我们通过（39）计算每个$\rho_k \in S_k$ 和 $f_k^d \in F_{k,\ d}$的最优数据传输速率 R_k^*，并选择使（38）中的代价函数最小的最优三元组 $T_k^* = (R_k^*,\ f_k^{d*},\ \rho_k^*)$。相反，为了评估第 k个用户设备（$d_k = 0$）上的本地学习任务的最小代价，我们只需穷举搜索满足条件的二元组 $P_k^* = (f_k^{d*},\ \rho_k^*)$，即

将在给定约束条件下优化准确率。最后，根据两种最优分配策略中哪一种能够保证最佳准确率，我们决定是否卸载第 k 个用户任务，即卸载（$d_k = 1$）或不卸载（$d_k = 0$），并分别采用相应的最优分配策略 T_k^* 或 P_k^*。

边缘服务器子问题：从边缘服务器的角度来看，该优化问题与 MU-MEDA 类似，仅成本函数略有不同，可表示为

$$\begin{aligned} \min_{\Phi_s} \quad & -\sum_{k=1}^{K}\sum_{i=1}^{L_k} \tau Q_{ki}^{ES} J_{ki}^{s} f_{ki}^{s} + \eta O \kappa \tau f_s^3 \\ \text{s.t.} \quad & 0 \le f_{ki}^{s}(t) \le \min\left(f_s, \frac{Q_{ki}^{ES}}{\tau J_{ki}^{s}}\right), \\ & \sum_{k=1}^{K}\sum_{i=1}^{L} f_{ki}^{s} \le f_s, \quad f_s \in \mathcal{F}_s, \qquad \forall k, t \end{aligned} \tag{40}$$

其中 $\Phi_s = [f_{ki}^s, f_s]$，并且可以采用与 MU-MEDA 公式相同的方法求解。

V. 仿真结果

本节展示我们通过所提出并求解的两种优化策略获得的仿真结果。表 II–III 列出了准确率 $G_k(\rho)$、第 k 个用户设备决定卸载分类任务时每个时钟周期内可压缩（并通过 JPEG 压缩）的数据单元 $J_k^d(\rho)$，以及同一用户设备每个时钟周期内可在本地压缩并分类的数据单元 $J_k^L(\rho)$。表 IV 列出了边缘服务器每个时钟周期内可分类的数据单元 $J_s(\rho)$，以及使用 JPEG 时短 CE 和 Deep-CE 共有的图像大小 $M(\rho)$ 和平均比特/像素数 $N(\rho)$。

表 I：LUT 参数

表 II：Deep-CE

ρ	$G(\rho)$ [%]	$J_k^d(\rho)$ $[\frac{DU}{C}]$	$J_k^L(\rho)$ $[\frac{DU}{C}]$
2	97.3	1.44×10^{-7}	8.35×10^{-8}
4	96.5	1.26×10^{-7}	9.04×10^{-8}
8	93.4	1.16×10^{-7}	8.90×10^{-8}
16	91.8	1.07×10^{-7}	8.73×10^{-8}
32	83.0	1.35×10^{-7}	1.06×10^{-7}
64	67.0	1.32×10^{-7}	1.09×10^{-7}

我们假设信道为平坦衰落信道，其统计特性基于 *Clarke* 自相关函数[48]。我们考虑了两种运行场景，如表 V 所示，并相应地将时隙时长设为$\tau = 50ms$，这与信道相干时间相对应。参数 σ_0^2 表示无线信道路径损耗功率

表 III：Short-CE

ρ	$G(\rho)$ [%]	$J_d(\rho)$ $[\frac{DU}{C}]$	$J_k^L(\rho)$ $[\frac{DU}{C}]$
2	97.3	1.44×10^{-7}	8.35×10^{-8}
4	95.8	1.68×10^{-7}	1.10×10^{-7}
8	91.5	1.88×10^{-7}	1.26×10^{-7}
16	91.3	1.95×10^{-7}	1.38×10^{-7}
32	77	2.25×10^{-7}	1.55×10^{-7}
64	50.0	2.25×10^{-7}	1.65×10^{-7}

表 IV：常用参数

ρ	$M(\rho)$ $[px]$	$N(\rho)$ $[\frac{bits}{px}]$	$J_s(\rho)$ $[\frac{DU}{C}]$
2	128x128x3	1.08	1.2×10^{-7}
4	64x64x3	2.27	2.17×10^{-7}
8	32x32x3	4.72	2.87×10^{-7}
16	16x16x3	9.06	3.57×10^{-7}
32	8x8x3	8	5×10^{-7}
64	4x4x3	8	6.25×10^{-7}

表 V：信道类型

信道类型	D $[m]$	B $[kHz]$	f_0 $[GHz]$	σ_0^2
A	50	2500	6	1.06×10^{-10}
B	500	2500	9	2.72×10^{-14}

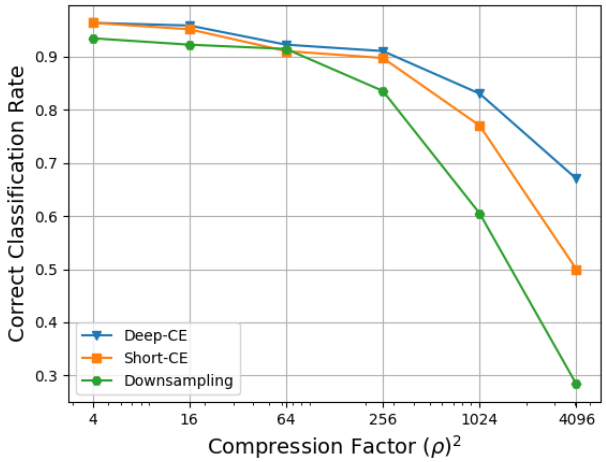

图3：分类准确率比较

路径损耗，计算时采用了*Alpha-Beta-Gamma* 模型 [49]。在第一组仿真中，我们考虑了一个有 $K = 5$ 个用户设备连接到网络的场景。虽然这并非严格必要，但我们假设所有用户设备的设备共用同一组计算频率 $\mathcal{F}_d = \{0.1, 0.2, \ldots, 0.9, 1\} \times 1.4\,GHz$，而服务器的计算频率组为 $\mathcal{F}_s = \{0.1, 0.2, \ldots, 0.9, 1\} \times 4.5\,GHz$。最后，为简便起见，我们假设所有用户设备和边缘服务器的有效开关电容均为$\kappa = 1.097 \times 10^{-27}[\frac{s}{cycles}]^3$。需要指出的是，所有仿真结果均是在所测试策略 [40]收敛后获得的。

A. 目标导向压缩结果

为简化起见，所有用户设备都被分配了相同的图像分类任务，该任务基于德国交通标志识别基准（GTSRB）[50] 数据集。该数据集包含1213张德国道路交通标志图片，分为43个不同类别。数据集被划分为80%的训练集（由970张图像组成）和20%的测试集（由243张图像组成）。在数据加载阶段，所有图像都被归一化为256x256的尺寸，并转换为3通道图像（每个RGB颜色对应一个通道），因此每个数据单元的初始尺寸为256x256x3。尽管这并非严格必要，但我们假设所有用户设备共享同一组CE-CC分类网络，例如，压缩因子 ρ_k 取值自同一固定集合 $\mathcal{S} = \{2, 4, 8, 16, 32, 64\}$。为了深入了解所提出的资源管理方法所取得的性能，

我们认为，有必要在图 3 中展示与不同压缩架构相关的测试集平均准确率：i) Deep-CE，ii) Short-CE，iii) 带抗混叠预滤波的下采样。正如预期的那样，对于所有模型，准确率 $G(\rho)$ 随压缩因子的增大而单调下降。Deep-CE 始终表现最佳，不过在压缩因子较低（最高为 16）时，与 Short-CE 的差异几乎可以忽略。相比之下，在压缩因子最高（即 32、64）时，使用 Deep-CE 的优势十分明显。当压缩因子为$\rho = 64$ 时，我们得到尺寸为 4x4x3=48 像素的输出张量：尽管这一尺寸的（伪）图像显然经历了大幅变换，Deep-CE 仍可使边缘服务器的 CC 以 67% 的准确率对其进行分类；对于一个包含 43 个类别的分类任务而言，这仍是相当出色的表现。相反，在这一压缩因子下，下采样策略和 Short-CE 都无法实现有意义的分类。提高 Deep-CE 准确率所付出的代价，是计算能耗和处理时延的增加（如表 II–III 所总结），我们通过资源管理策略对其进行权衡。

B. mu-MEDA 结果

首先，我们在信道场景 B中测试了 mu-MEDA 策略，将 CE（浅层和深层）与下采样压缩策略进行比较，结果如表 V 所示。我们为所有用户设备设置了相同的时延 约束_ $D_k^{avg} = Q_k^{avg}/A_k = 0.20\ s$。我们考虑了一个任务到达过程 $A_k = 2DU/slot$，并强制用户设备始终将分类任务卸载至边缘服务器，不采用任何机会式策略（即，$d_k(t) = 1,\ \forall k,\ t$）。

图 4 和图 5 中的每条权衡曲线都对应于不同的准确率约束，而它们均满足相同的时延约束，该约束在图中以一条水平虚线突出显示。每条曲线都是通过针对（23）式中的多个不同权衡参数值 V，评估资源优化问题的解（收敛时）而获得的。具体而言，随着 V 的增大，最终得到的解具有更低的能耗和更高的时延；如图中的黑色箭头所示，我们沿着权衡曲线从右下角移动到左上角，这些位置对应于可行区域边界上所期望的最优解。图 4 表明，从 UE 的角度来看，采用 CE 压缩策略具有明显优势，因为最终得到的解具有更低的（计算和传输）能耗，同时满足相同的时延和准确率约束。这取决于这样一个事实：信道- B 具有极大的衰减。因此，由于 CE 压缩策略能够通过传输比传统下采样更小的数据单元来满足相同的准确率约束，因此可以显著降低传输能耗，而无需在 UE 端为基于 CE 的压缩消耗过多额外的计算能耗。事实上，如图 5 所示，所提出的动态、目标导向压缩策略还降低了 ES 的计算能耗。实际上，传统的.....

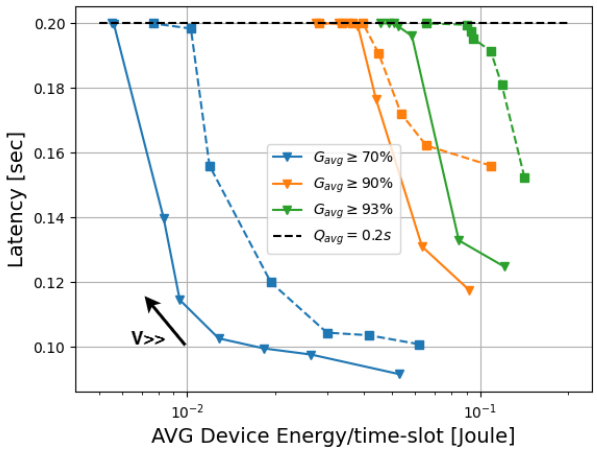

图 4：用户设备的能耗/时延权衡。交叉熵（实线）与下采样（虚线）。

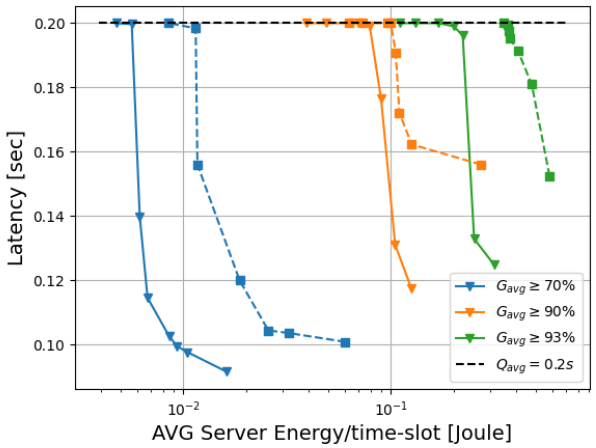

图 5：边缘服务器的能耗/时延权衡。交叉熵（实线）与下采样（虚线）。

从计算和能耗角度来看，较小数据单元的分类成本更低。

UE	通道类型	$\kappa[\frac{s}{cycles}]^3$
0	**A**	$10 \times \kappa_0$
1	**A**	$20 \times \kappa_0$
2	**B**	$30 \times \kappa_0$

表 VI：每个用户设备的仿真场景

C. 机会式卸载

我们将前一种场景与机会式卸载策略进行了比较：在前一种场景中，用户设备始终将决策任务卸载到边缘服务器；而在机会式卸载策略中，用户设备也可以决定在本地执行分类。两种策略均采用相同的交叉熵-级联分类架构。具体而言，五台用户设备中有两台通过表 V 的场景 A 中的信道连接到边缘服务器，其余设备则通过场景 B中的信道连接。机会式卸载策略最终实现了动态资源优化，与始终卸载 策略相比，用户设备能耗显著降低，同时仍满足准确率和时延约束。如图 6–7 所示，所有实线都明显位于虚线左侧，即能耗更低。虚线表示纯卸载策略的结果

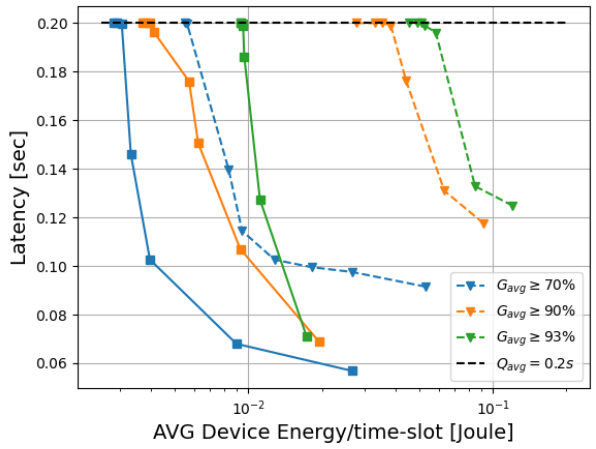

图 6：用户设备的能耗/时延权衡。机会式卸载（实线）与仅卸载策略（虚线）的比较。

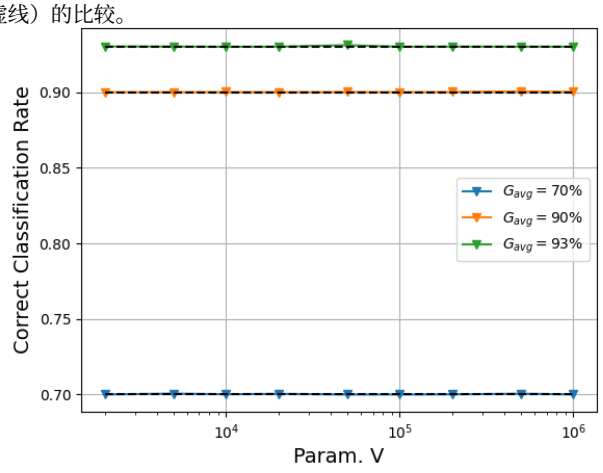

图 7：收敛时采用机会式卸载的平均准确率与 V

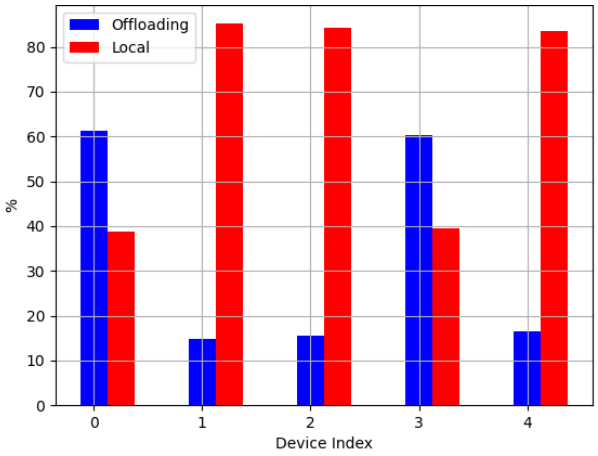

图 8：卸载百分比（$G_k^{avg} = 70\% \forall k, V = 1 \times 10^6$）。

曲线。图 8 展示了在最低准确率约束为 $G_{avg} = 70\%$、权衡参数为 $V = 1 \times 10^6$ 时各用户设备的卸载决策直方图。不出所料，由于用户设备 0 和用户设备 3 的信道条件较好，它们决定卸载的频率高于其他设备；对于其他设备而言，信道 B 需要更高的发射功率才能分配速率

用户设备有时也可能无法同时满足准确率约束、时延约束，或两者都无法满足。

D. 与静态分配策略的比较

所提方法的一项关键优势在于联合动态 优化传输和计算资源，同时动态选择用于执行任务的最优分类架构。因此，我们将所提出的多用户优化策略与以下方法进行比较：

- 一种固定准确率优化策略：固定单一交叉熵-级联分类架构，同时优化用户设备侧的计算资源和传输资源。该方法与 [6] 中提出的方法非常相似。

- 一种混合静态/动态优化策略，受 [51]，启发，我们根据平均信道条件固定传输速率 R ，同时动态优化交叉熵-级联架构以及用户设备的计算资源。传输速率 R 固定为能够保证用户设备队列稳定的最小速率。该速率可根据平坦衰落瑞利信道的容量计算得出（ [52] 中的公式 (9)），并且也由此确定传输功率。

在这种情况下，我们考虑了一个包含 $K = 3$ 个用户设备的场景，每个用户设备具有不同的信道条件和计算效率，如表 VI 所示。我们设置到达任务为 $A = 2DU/slot$，并为所有用户设备设定相同的准确率和时延约束，分别为 $G_k^{avg} = 92\%$ 和 $D_k^{avg} = 0.2s$。因此，对于固定准确率优化策略，我们将 Short-CE 的 $\rho_k = 8$ 作为唯一学习模型；根据表 III，该模型能够以适中的计算能耗达到所需的平均分类性能。图 9 显示，采用完全动态优化策略可得到用户设备能耗更低的解决方案。正如预期的那样，UE-0 和 UE-2 的能耗分别最低和最高，这与表 VI 所概述的计算条件和信道条件相符。显然，对于所有用户设备，我们的优化策略都能实现最低能耗，从而证实了联合动态优化传输/计算资源以及学习架构（即所采用的 CE-CC 对）的有效性，并根据系统瞬时条件进行调整。

E. mu-MADE 结果

我们测试了 mu-MADE 优化策略，考虑了一个包含 $K = 3$ 个用户设备的场景，每个用户设备都具有不同的信道和计算条件。具体而言，我们为边缘服务器设定了有效开关电容 κ_0=$1.097 \times 10^{-27}[\frac{s}{cycles}]^3$，并为用户设备设定了更高的值，以模拟较低的能效。用户设备能量约束设为 $E_k^{avg} = 128 \times 10^{-3} J$。表 VI 汇总了仿真中所考虑设备的不同条件，我们同时采用了 Deep-CE 和 Short-CE。值得注意的是，UE-0 具有

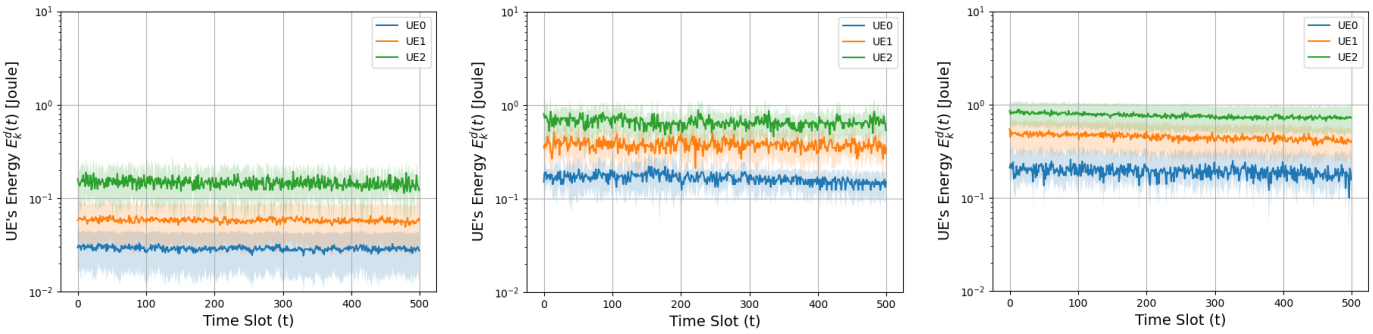

图 9：动态优化（a）、固定准确率（b）和固定速率（c）下用户设备的瞬时能耗。

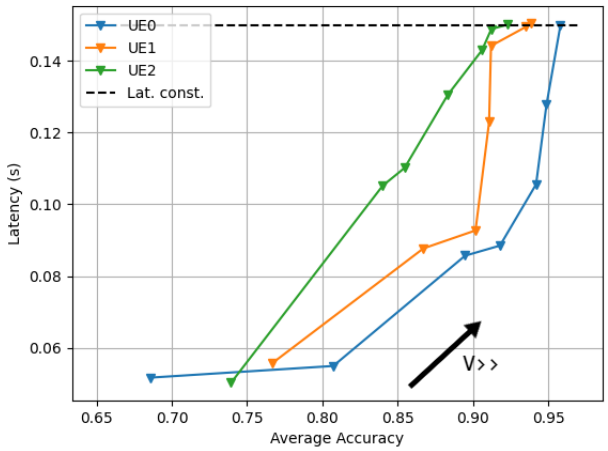

图 10：准确率与时延之间的权衡。

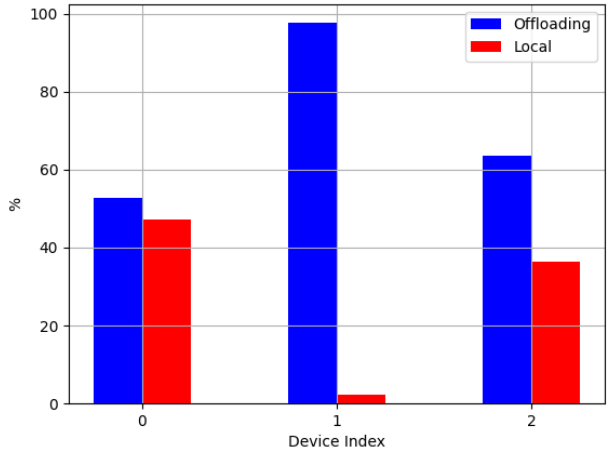

图 11：卸载直方图（$V = 1 \times 10^5$）。

良好的信道条件和计算效率：这意味着它在机会式卸载管理方面拥有最大的灵活性。UE-1 的信道条件与 UE-0 相同，但计算效率较低；而 UE-2 的信道条件较差，计算能效也较低。

图 10 中所示的曲线体现了准确率-时延权衡：通过增大式 (36) 中的参数 V，我们最终得到准确率和时延更高的解，沿曲线从左下角移动到右上角，其中我们在决策区域的边界处得到理想的最优解。具体而言，图 10 显示，UE-0（即计算能力和信道条件最佳的用户设备）获得了最高准确率，同时大幅满足了时延约束。UE-1 和 UE-2 也呈现出类似的行为，不过 UE-2 的时延更高（即该设备所处的条件最差）。最后，我们在图 11 中报告了各用户设备的卸载决策直方图。由于信道条件良好且计算能耗效率高，UE-0 的情况较为均衡，因为它在选择是否卸载计算任务方面具有最大的灵活性。另一方面，UE-1 主要进行卸载，因为在衰减较低的信道中传输数据单元，可以减轻计算能耗效率低下所带来的负担。最后，尽管 UE-2 的信道条件差得多，但由于其计算效率低得多，它卸载的数据单元比 UE-0 多。

VI. 结论与未来研究方向

在这项工作中，我们实现了一种基于交叉熵的目标导向压缩架构，并采用两种不同的动态优化策略，在多用户场景中实现能耗最小化或学习准确率最大化；在该场景中，用户设备可以机会性地决定是否以及何时将计算任务卸载至边缘服务器。大量仿真结果证实了所提出方法在不同场景下的有效性和灵活性。不过，我们指出，所提出的目标导向通信架构及其相关资源管理策略，也可以采用分类或面向学习的压缩策略，而这些策略可能不同于本文提出的基于交叉熵的方案。未来的研究方向包括扩展到多服务器场景、协作学习任务（例如联邦学习），以及明确考虑每台用户设备的电池电量；这些设备还可以配备能量采集机制或电池充电计划。

附录 A
MU-MEDA 的数学推导

[40] 中的两个引理有助于解决所提出的资源优化策略。

引理 *A.1*：给定一个按$X(t+1) = max(0, X(t) + x(t+1) - x)$ 演化的队列，通过定义 $\Delta_x =$

$\frac{X(t+1)^2 - X(t)^2}{2}$，始终有 $\Delta_x \leq \frac{(x(t+1)^2 - x(t)^2)}{2} + X(t)x(t+1) - X(t)x$。

引理 *A.2*：以下不等式成立：

$$(\max(0, Q-b)+A)^2 \leq Q^2 + A^2 + b^2 + 2Q(A-b).$$

利用引理 A.1，并回顾给定 $x \in \mathbb{R}^k$ 时，对于时延虚拟队列 $Z_k(t)$，$(\sum_{k=1}^K x_k)^2 \leq K\sum_{k=1}^K x_k^2$，我们有

$$\begin{aligned}\Delta_{z_k}(t) &\leq \frac{\mu_k^2(Q_k^{tot}(t+1) - Q_k^{avg})^2}{2} \\ &+ \mu_k Z_k(t)(Q_k^{tot}(t+1) - Q_k^{avg}) \\ &\leq \frac{\mu_k^2 L_k}{2}[Q_k^{UE}(t+1)^2 + \sum_{i=1}^{L_k} p_i Q_{ki}^{ES}(t+1)^2] \\ &+ \mu_k Z_k(t)(Q_k^{tot}(t+1) - Q_k^{avg}) + \frac{\mu_k^2 L_k}{2}(Q_k^{avg})^2,\end{aligned}$$

现在，回顾式 (8)、(9)，并利用引理 A.2，我们可以推导出以下不等式

$$\begin{aligned}\Delta_{z_k}(t) \leq \frac{\mu_k^2 L_k}{2}\{&Q_k^{UE}(t)^2 + M_k^{UE} + 2Q_k^{UE}(t)(A_k(t) \\ &- N_k^{UE}(t)) + \sum_{i=1}^{L_k}[p_i Q_{ki}^{ES}(t)^2 + M_k^{ES} \\ &+ 2p_i Q_{ki}^{ES}(t)(\hat{A}_{ki}(t) - N_{ki}^{ES}(t))]\} \\ &+ \mu_k Z_k(t)(Q_k^{tot}(t+1) - Q_k^{avg}) + \frac{\mu_k^2 L_k}{2}(Q_k^{avg})^2\end{aligned}$$

其中 $\hat{A}_{ki}(t) = \mathbb{1}\{\rho_k(t) = s_{ki}\}N_k^{UE}(t)$，$M_k^{UE} = A_{k,\,max}^2 + N_{kdev,\,max}^2$ 以及 $M_{ik}^{ES} = A_{ki,\,max}^2 + N_{ki,\,max}^2$。

[9] 中给出的相同推导过程可以应用于准确率虚拟队列，从而得到 $\Delta_{y_k}(t)$ 的一个上界。将推导出的瞬时上界汇总后，我们最终得到第 IV-A 节所 ներկայաց出的优化问题。

参考文献

[1] W. Jiang, B. Han 等，“迈向 6G：一项全面综述”，*IEEE* 通信学会开放期刊，第 2 卷，366，2021 年，334 页。– .[2] Z. Zhou, X. Chen 等，“边缘智能：借助边缘计算铺平人工智能的最后一公里”，*IEEE* 会刊，第 107 卷，第 8 期，第 1738–1762 页，2019 年。[3] J. Park, S. Samarakoon 等，“边缘侧的无线网络智能”，*IEEE* 会刊，第 107 卷，第 11 期，第 2204–2239 页，2019 年。[4] S. Wang, T. Tuor 等，“当边缘遇上学习：资源受限分布式机器学习的自适应控制”，载于*IEEE INFOCOM 2018* 会议论文集，2018 年，第 63–71 页。[5] G. Zhu, D. Liu 等，“迈向智能边缘：无线通信与机器学习相遇”，*IEEE* 通信杂志，第 58 卷，第 1 期，第 19–25 页，2020 年。[6] M. Merluzzi, P. Di Lorenzo 等，“无线边缘机器学习：资源分配与权衡”，*IEEE Access*，第 9 卷，第 45 377–45 398 页，2021 年。[7] M. Giordani, M. Polese 等，“迈向 6G 网络：用例与技术”，*IEEE* 通信杂志，第 58 卷，第 3 期，第 55–61 页，2020 年。[8] E. C. Strinati 和 S. Barbarossa，“6G 网络：超越香农，迈向语义与目标导向通信”，计算机网络，第 190 卷，文章编号 107930，2021 年。[9] M. Merluzzi, P. Di Lorenzo 等，“具备时延和准确率保证的无线边缘机器学习动态资源分配”，载于*IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2020 年，第 9036–9040 页。

[10] M. Merluzzi、C. Battiloro 等，“具有可靠性保证的无线边缘能效分类”，载于《2022 IEEE 国际通信会议研讨会（ICC Workshops）》，2022 年，第 109–114 页。[11] M. Merluzzi、A. Martino 等，“边缘动态集成推理”，载于《IEEE 全球通信会议》。IEEE，2021 年，第 1–6 页。[12] N. Tishby、F. C. Pereira 等，“信息瓶颈方法”，预印本 physics/0004057，2000 年。[13] Z. Goldfeld 和 Y. Polyanskiy，“信息瓶颈问题及其在机器学习中的应用”，《IEEE 信息论选定领域期刊》，第 1 卷，第 1 期，第 19–38 页，2020 年。[14] F. Pezone 等，“基于信息瓶颈的面向目标通信与边缘学习”，载于《ICASSP 2022-2022 IEEE 国际声学、语音与信号处理会议（ICASSP）》。IEEE，2022 年，第 8832–8836 页。[15] G. Chechik、A. Globerson 等，“高斯变量的信息瓶颈”，《机器学习研究期刊》，第 6 卷，第 1 期，2005 年。[16] J. Shao、Y. Mao 等，“面向边缘推理的任务导向通信学习：一种信息瓶颈方法”，《IEEE 通信选定领域期刊》，第 40 卷，第 1 期，第 197–211 页，2022 年。[17] ——，“面向多设备协作边缘推理的任务导向通信”，《IEEE 无线通信汇刊》，2022 年。[18] E. Bourtsoulatze、D. B. Kurka 等，“用于无线图像传输的深度联合源信道编码”，《IEEE 认知通信与网络汇刊》，第 5 卷，第 3 期，第 567–579 页，2019 年。[19] C.-H. Lee、J.-W. Lin 等，“面向物联网的深度学习构建联合传输-识别”，《IEEE Access》，第 7 卷，第 76 547–76 561 页，2019 年。[20] M. Jankowski、D. G¨und¨uz 等，“边缘无线图像检索”，《IEEE 通信选定领域期刊》，第 39 卷，第 1 期，第 89–100 页，2020 年。[21] T.-Y. Tung、D. B. Kurka 等，“Deepjscc-q：星座约束深度联合源信道编码”，预印本 arXiv:2206.08100，2022 年。[22] M. Yang、C. Bian 等，“OFDM 引导的适用于无线多径衰落信道的深度联合源信道编码”，《IEEE 认知通信与网络汇刊》，2022 年。[23] J. Dai、S. Wang 等，“用于语义通信的非线性变换源信道编码”，《IEEE 通信选定领域期刊》，2022 年。[24] J. Ball´e、P. A. Chou 等，“非线性变换编码”，《IEEE 信号处理选定主题期刊》，第 15 卷，353，2020 2，339，第 期，页 – 。[25] M. K. Farshbafan、W. Saad 等，“面向目标语义通信的通用语言：一种课程学习框架”，载于《ICC 2022-IEEE 国际通信会议》。IEEE，2022 年，第 1710–1715 页。[26] ——，“使用通用语言的面向目标语义通信课程学习”，预印本 arXiv:2204.10429，2022 年。[27] H. Xie、Z. Qin 等，“面向任务的多用户语义通信”，《IEEE 通信选定领域期刊》，第 40 卷，第 9 期，第 2584–2597 页，2022 年。[28] X. Peng、Z. Qin 等，“一种稳健的、基于深度学习的文本语义通信系统”，预印本 arXiv:2206.02596，2022 年。[29] P. Jiang、C.-K. Wen 等，“用于句子语义传输的深度源信道编码与 HARQ”，《IEEE 通信汇刊》，2022 年。[30] X. Kang、B. Song 等，“面向无人机系统场景分类的任务导向图像传输”，《IEEE 通信汇刊》，第 70 卷，第 8 期，第 5181–5192 页，2022 年。[31] H. Xie、Z. Qin 等，“面向 VQA 任务的任务导向多用户语义通信”，《IEEE 无线通信快报》，第 11 卷，第 3 期，第 553–557 页，2021 年。[32] Z. Weng、Z. Qin 等，“用于语音识别的语义通信”，预印本 arXiv:2107.11190，2021 年。[33] F. Binucci、P. Banelli 等，“面向目标通信的边缘推理自适应资源优化”，《EURASIP 信号处理进展期刊》，第 2022 卷，第 1 期，第 1–34 页，2022 年。[34] ——，“无线边缘多用户面向目标通信的动态资源分配”，载于《2022 年第 30 届欧洲信号处理会议（EUSIPCO）》。IEEE，2022 年，第 697–701 页。[35] G. Chechik、A. Globerson 等，“高斯变量的信息瓶颈”，《机器学习研究期刊》，第 6 卷，第 1 期，2005 年。[36] K. G. Larkin，“香农信息论思考：探寻图像的自然信息熵”，预印本 arXiv:1609.01117，2016 年。

[37] M. Boudiaf、J. Rony 等，“度量学习的统一互信息视角：交叉熵与成对损失之比较”，载于 欧洲计算机视觉会议。Springer，2020 年，第 548–564 页。[38] L. Le、A. Patterson 等，“监督式自编码器：利用无监督正则化器提升泛化性能”，载于 神经信息处理系统进展，第 31 卷。Curran Associates, Inc.，2018 年。[39] S. Dodge 和 L. Karam，“理解图像质量如何影响深度神经网络”，载于 *2016* 年第八届多媒体体验质量国际会议（*QoMEX*）。IEEE，2016 年，第 1–6 页。[40] M. J. Neely，“随机网络优化及其在通信与排队系统中的应用”，通信网络综合讲义，第 3 卷，第 1 期，第 1–211 页，2010 年。[41] J. D. Little，“排队公式的证明：L= λ w”，运筹学，第 9 卷，第 3 期，第 383–387 页，1961 年。[42] C. E. Shannon，“通信的数学理论”，贝尔系统技术期刊，第 27 卷，第 3 期，第 379–423 页，1948 年。[43] T. Burd 和 R. Brodersen，“便携式系统的处理器设计”，*VLSI* 信号处理期刊，第 13 卷，11 1996。[44] N. Tishby 和 N. Zaslavsky，“深度学习与信息瓶颈原理”，载于 *2015 IEEE* 信息论研讨会（*ITW*）。IEEE，2015 年，第 1–5 页。

[45] A. M. Saxe, Y. Bansal等，《深度学习的信息瓶颈理论》，统计力学杂志：理论与实验，第 2019 卷，第 12 期，第 124020 页，2019 年。[46] S. Boyd 和 L. Vandenberghe，《凸优化》。剑桥大学出版社，2004 年。[47] M. Assi 和 R. A. Haraty，《背包问题综述》，载于*2018* 年阿拉伯国际信息技术会议（*ACIT*），2018 年，第 1–6 页。[48] A. F. Molisch，《无线信道的统计描述》，2011 年。[49] S. Sun, T. S. Rappaport等，《5G 城市微蜂窝和宏蜂窝场景的传播路径损耗模型》，载于*2016* 年 *IEEE* 第 *83* 届车辆技术会议（*VTC Spring*）。IEEE，2016 年，第 1–6 页。[50] J. Stallkamp等，《德国交通标志识别基准：一项多类别分类竞赛》，载于*2011* 年国际神经网络联合会议。IEEE，2011 年，第 1453–1460 页。[51] C. Liu, C. Guo等，《面向任务通信的自适应语义压缩与资源分配》，预印本 *arXiv:2204 08910*，2022。。[52] J. Li, A. Bose等，《瑞利平坦衰落信道的容量》，载于第 *3* 届年度通信网络与服务研究会议（*CNSR'05*）。IEEE，2005 年，第 214–217 页。