计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330005
60(4): 786−796, 2023

算力网络支撑下的泛在化视频传输调度

张旭光　陈鸣锴　魏　昕
（南京邮电大学通信与信息工程学院　南京　210003）
（宽带无线通信与传感网技术教育部重点实验室（南京邮电大学）南京　210003）
（xwei@njupt.edu.cn）

Ubiquitous Video Transmission Scheduling Supported by Computing Power Network

Zhang Xuguang, Chen Mingkai, and Wei Xin
（*College of Telecommunications & Information Engineering*, *Nanjing University of Posts and Telecommunications*, *Nanjing* 210003）
（*Key Laboratory of Broadband Wireless Communication and Sensor Network Technology* (*Nanjing University of Posts and Telecommunications*), *Ministry of Education*, *Nanjing* 210003）

Abstract　The explosive growth of video data volume, the increasing diversity of video forms, and the ubiquity of video services are the three main characteristics of the development of current video communication technology. This fact will undoubtedly lead to two main problems. The first is that the traffic burden of the core network is difficult to offload, and the second is that, the video transmission conflict is difficult to coordinate. In order to alleviate these two problems, we propose a ubiquitous video scheduling scheme, supported by computing power networks. Specifically, we first propose a hierarchical video coding and decoding model to enhance the flexibility of video content deployment and transmission; Secondly, we propose a service-oriented good-put model with the consideration that the diversity transmission parameter constraints of different video services; Finally, with the support of computing power network, on the one hand, through task decomposition, the "fragmented" network resources can be utilized effectively, and on the other hand, through global detection and real-time perception of network status, accurate deployment of video content and efficient scheduling of network resources can be achieved. Simulation experimental results verify the effectiveness of the proposed scheme in terms of core network traffic offloading, good-put improvement, and network resource utilization rate improvement.

Key words　computing power network; ubiquitous video; video scheduling; good-put; content distribution

摘　要　视频数据量的爆炸式增长、视频形式愈加多样、视频业务的泛在化是当前视频通信技术发展的三大特点，这无疑会导致核心网过载和视频传输调度难的问题. 为了缓解这些问题，提出一种算力网络支撑下的泛在化视频传输调度方案. 具体地，首先提出一种视频层次化编解码模型提升视频内容部署和传输的灵活性；其次，考虑视频业务的差异化指标约束，提出面向业务的有效吞吐量模型；最后，借助算力网络的支撑，一方面通过任务分解来有效利用"碎片化"的网络资源，另一方面通过对网络状态的全局检测和实时感知，实现视频内容的精准部署和网络资源的高效调配. 仿真实验验证了所提方案在核心网流量卸

收稿日期：2023-01-03；**修回日期**：2023-02-01
基金项目：国家自然科学基金项目（62071254, 62231017, 62001246）；江苏高校优势学科建设工程项目；南京邮电大学引进人才自然科学研究启动基金（NY222035）；中国博士后科学基金资助项目（2022M721694）；江苏省卓越博士后计划（2022ZB398）
This work was supported by the National Natural Science Foundation of China (62071254, 62231017, 62001246), the Priority Academic Program Development of Jiangsu Higher Education Institutions, the Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (NY222035), the Project Funded by China Postdoctoral Science Foundation (2022M721694), and the Jiangsu Funding Program for Excellent Postdoctoral Talent (2022ZB398).
通信作者：魏昕（xwei@njupt.edu.cn）

载、有效吞吐量提升、网络资源利用率提升方面的有效性.

关键词 算力网络；泛在化视频；视频调度；有效吞吐量；内容分发

中图法分类号 TP391

随着电子技术、网络技术以及信息技术的发展，视频已成为网络承载的主要流量[1]. 尤其在近几年，受新型冠状病毒肺炎（corona virus disease 2019, COVID-19）疫情的影响，线上办公、视频会议已成为人们工作的常态. 短视频业务、视频广告、视频游戏等也已渗透入人们生活的方方面面. 正在涌现的自动驾驶、远程手术、虚拟现实（virtual reality, VR）、增强现实（augmented reality, AR）、混合现实（mixed reality, MR）等应用均离不开高质量视频通信的技术加持. 毫不夸张地说，5G 赋能下的视频通信技术已经改变了人类社会.

现如今，视频通信的发展呈现 3 个特点：

1）视频质量大幅提升，视频数据量呈指数级增长. 720p, 1080p, 4K 分辨率的视频服务已经普及，8K 及以上分辨率视频通信技术也早已实现. 视频帧率也从满足人类最基本视觉暂留效应的 30 fps 提升至 60 fps, 90 fps 乃至 120 fps, 以提供更加优质的视频交互体验，如大型视频游戏、多视点赛事直播、VR、AR 交互等. 据统计，一场多视点的体育赛事直播每小时会产生超过 7 GB 的数据量，VR 或 AR 应用所需的带宽预计超过 25 Mbps, 8K UHD 超高清流媒体对单个用户的容量要求更是高达 100 Mbps[2]. 如此庞大的视频数据量无疑会急剧加重网络的传输负担，仅通过传统方式提供视频服务将使核心网无法承载.

2）视频服务更加多元化，视频形式更加多样化. 除了传统不同质量的视频并存于网络之外，新兴的 360°全景视频[3]、多视点视频[4] 涌现，并可用于支持 AR, VR, MR 类应用. 此外，同一视频内容支持多种视频呈现形式，以满足不同用户的不同服务需求. 例如一场体育赛事直播，既要向上支持 360°全景、多视点等形式的沉浸式体验服务，又要向下兼容不同分辨率的视频流媒体服务，以适应不同用户的差异化观看场景.

3）视频服务趋于泛在化，视频内容提供商趋向“草根”化. 网络技术的发展和智能终端的普及，推动了多媒体业务的广泛推广. 近年来，视频已成为人们获取信息的主要载体. 尤其是抖音、腾讯、Instagram 等短视频应用的高度流行[5]，为人们提供随时随地视频服务接入的同时，用户本身也成为了视频内容的创作者和提供商. 这一方面导致移动数据流量的爆炸式增长，给底层移动网络带来了巨大的压力. 另一方面，这对传统集中式的网络构架提出了挑战. 如果集中式地处理泛在化的视频内容，不仅会导致庞大的用户接入量难以协调的问题，而且会引入额外的回程延时，降低视频应用的服务质量.

在此激励下，算力网络（computing power network）[6] 应运而生. 为应对无处不在的网络连接，算力网络将动态分布的计算与存储资源互联，通过网络、存储、算力等多维度资源的统一协同调度，使海量应用能够按需、实时调用泛在分布的计算资源，实现连接和算力在网络的全局优化，提供一致的用户体验[7]. 总体而言，算力网络旨在将网络中动态分布的计算、存储、通信等网络资源统一管理并充分利用，将服务的提供方式由集中式的统一供应转换为分布式的云、边、端协作. 这一方面可以提升网络可用资源的总量，卸载核心网的负载，另一方面可以使“服务提供商”更靠近用户，降低服务延时. 因此算力网络适用于对网络资源需求大的业务，如超算和大型渲染、人工智能、自动驾驶、和区块链应用等. 尤其是高质量、泛在化的视频业务，如智能安防、云 VR/AR、超高清视频、云游戏、元宇宙等.

然而，将算力网络应用于泛在化视频业务中仍然面临挑战：算力网络试图开发动态分布的网络资源以提升网络性能，然而“碎片化”是这类网络资源的主要特征之一. 例如，智能终端和多接入边缘计算（multi-access edge computing, MEC）服务器中的计算和存储资源，虽分布广泛，但其密集程度仍无法与云中心相比. 当处理资源密集型的视频业务时，资源分布“碎片化”与资源需求“密集型”之间的矛盾存在如何协调的问题. 为了解决这个问题，本文提出一种算力网络支撑下的视频传输调度方案，如图 1 所示. 具体包含 3 个主要贡献：

1）提出了一种层次化视频编解码模型，通过视频主成分提取和消除，提供一种独立-可伸缩的视频解构和重构方案，便于在算力网络中灵活传输；

2）基于内容分布的全局观测和最大化内容部署的增益度，设计视频内容的按需部署策略，提升视频内容的命中率从而提升碎片化网络资源的利用率；

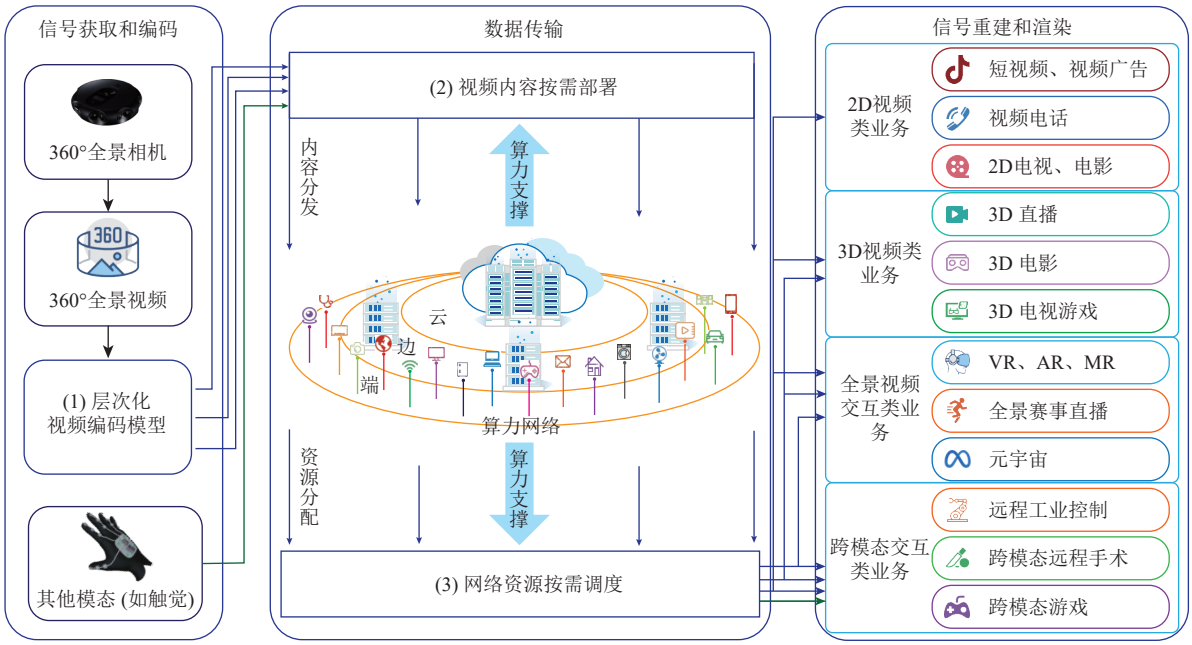

Fig. 1 Ubiquitous video communication model supported by the computing power network

图 1 算力网络支撑下的泛在化视频通信模型

3) 基于业务的差异化指标约束, 设计传输链路的有效吞吐量模型, 并设计基于有效吞吐量的链路选择算法, 提升视频传输效率.

1 相关工作

无线网络中的视频传输需要传输码流具有对无线信道的适应性. 由于易于部署和扩展, 目前流行的流媒体业务大都基于超文本传输协议(Hyper Text Transfer Protocol, HTTP)的自适应流媒体[8](HTTP adaptive streaming, HAS)进行视频传输. HAS 的基本思想是通过存储若干个不同质量的视频版本来适应网络环境. 客户端依据网络环境和自身需求, 自主选择合适的视频码率版本进行传输, 从而达到网络自适应的目的. 诸如 Adobe 公司的 Adobe HTTP Dynamic Streaming、苹果公司的 Apple HTTP Live Streaming、微软公司的 Microsoft Smooth Streaming, 以及 HTTP 动态自适应流(dynamic adaptive streaming over HTTP, DASH)都是流行的 HAS 协议. 以 DASH[9] 为例, 它支持不可伸缩的或可伸缩的视频表征方法. 不可伸缩的方法即预先存储或实时地将视频转码为多种质量的视频码流, 然后根据需要发送合适的版本给客户端. 显然, 这种方法依赖视频转码或者视频联播技术的支持. 实时的视频转码需要视频服务器强大的计算能力支持, 而预先存储多个视频质量版本的视频联播技术显然会耗费视频服务器大量的存储空间. 因此这种方案的部署依赖强大的计算、存储中心支持, 无法从根本上缓解算力网络资源“碎片化”与视频业务资源需求“密集型”之间的矛盾.

从码流结构上入手, 将资源需求“密集型”的视频业务解构为多任务协作是一种提升“碎片化”网络资源的有效方法. 例如通过利用可伸缩视频编码(scalable video coding, SVC)技术来适配网络资源和业务需求, 便可提升视频传输的灵活性. DASH 中的可伸缩方法[10] 即为一种典型的应用. 该方法通过预先将视频编码为若干个视频层, 视频服务器依据网络条件发送给客户端不同数量的视频层数据以实现视频恢复质量的可伸缩性和对网络的适应性. 由于 SVC 可以通过分层编码实现一次编码、多种解码, 从而无需服务器存储多个视频质量副本或转码, 很大程度上降低了承担视频服务器的负担(例如一些智能终端设备就可以为其它设备提供视频内容). 此外, 不仅仅是 DASH 方案支持可伸缩的视频表征, 基于 SVC 的可伸缩视频流在异构网络(heterogeneous networks, HetNets)[11-13]、设备到设备(device-to-device, D2D)网络[14-16]、MEC 支持的蜂窝网络[17-18] 中的传输问题也引起了广泛关注. 此类研究中, 为了发挥 SVC 在灵活传输中的优势, 一个核心问题是如何部署视频的缓存副本提升用户设备请求视频的命中概率. 根据相关研究, 视频的受欢迎程度通常在很大程度上影响视频的部署, 例如基于视频内容的受欢迎程度通常呈现 Zipf[19] 分布的现实, 在网络中缓存 Zipf

偏度参数较高(代表了视频请求的集中程度高)的视频内容，会使得视频内容的命中率得以提升. 但这种方法同时也会导致无效缓存的问题，亟待进一步研究解决.

借鉴以上研究的经验，本文着重研究算力网络中的泛在化视频调度问题. 但不得不解决 2 个问题：1)SVC 的天生属性使得其上层数据的解码严重依赖底层数据，这种设计的初衷是提升视频编码的效率，但却也使得其码流结构不够灵活，将其直接应用于资源碎片化分布的算力网络中时，其性能优势难免受到抑制；2)基于 Zipf 分布的视频内容部署是从概率角度缓存可能会被请求的视频内容，难免与用户的实际请求不符，造成无效缓存的问题，进而浪费网络资源. 鉴于此，本文从层次化的视频编码入手，解决视频流对算力网络的适应性问题，促进网络资源的灵活调度. 在此基础之上，以视频内容的按需部署和网络资源的按需调度，提升视频业务的适应性和“碎片化”网络资源的利用率.

综上所述，泛在化视频业务的高效调度一方面需要通过将资源“密集型”的视频传输任务分解，从而开发算力网络中碎片化的网络资源，缓解爆炸式增长的视频流量对核心网的冲击；另一方面需要借助算力网络提供的算力支撑，实现视频内容的精准部署和网络资源的高效调配，在确保视频业务服务质量的同时，兼顾视频业务的多样性特点和无线网络的动态性特征.

2 系统模型

本节主要介绍算力网络支撑下的泛在化视频通信模型，主要包括视频的层次化编解码模型、面向业务的有效吞吐量模型.

2.1 视频的层次化编解码模型

为了应对泛视频业务的多样性、网络资源的碎片化和无线网络的动态性，本文提出一种层次化的视频编码模型如图 2 所示[20]. 以空间域为例，首先，视频信号被下采样为多路子信号，以提高码流的适应性，降低基本解码门限；其次，将下采样的若干路子信号中相关的“主成分”提取成一路主信号，并将该主成分从各个子路中去除，以避免各个子路相互依赖，提升传输和解码的灵活性；再次，按照传统视频编码的方法，对主路信号及各个子路信号进行时—空相关性编码和熵编码，以去除视频信号的时—空冗余和统计冗余. 其中，各子路信号可参考主路信号进行参考编码，进一步去除层间冗余. 最后，经过本文提出的编码模型，视频信号被编码成一路主描述层和若干路增强描述层. 主描述层可以以基本的视频质量单独解码，增强描述层则用于提升视频质量，且解码的增强描述层数越多，视频质量越高. 但不同于 SVC 或 H.265 高效可伸缩视频编码(scalable high-efficiency video coding, SHVC)的是，经过本文提出的编码模型生成的增强描述层信号只参考主描述解码而不相互依赖，即解码质量只与解码层数有关，与顺序无关. 且不同于多描述编码(multiple description coding, MDC)，各个增强描述之间的相关成分被提取到主描述中，因此视频编码效率得以提升.

以空间域为例阐述本文提出的层次化视频编码模型具体技术实现：

1)将原始视频信号以 P^o 表示，其每一帧的分辨率为 (X, Y)，总帧数为 K.

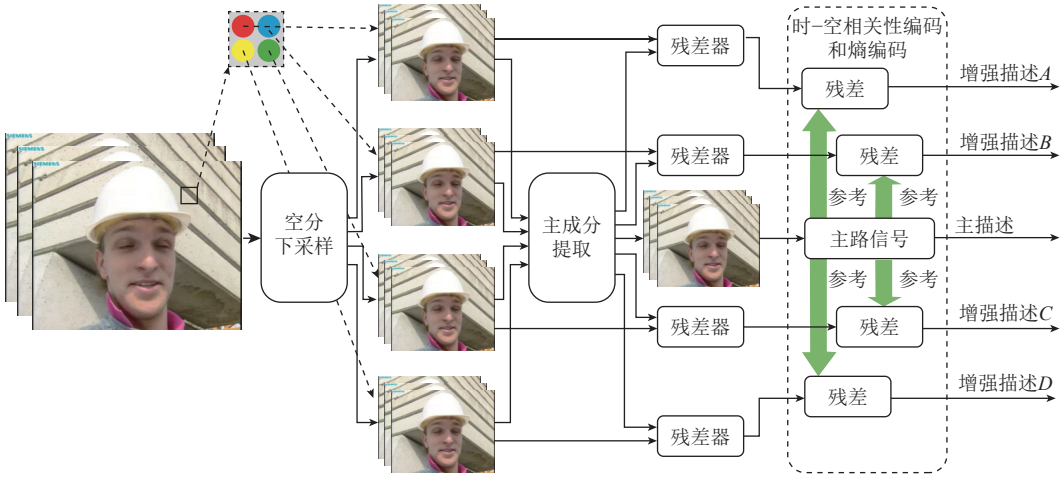

Fig. 2 Hierarchical coding model for ubiquitous video services (take spatial domain as an example)

图 2 适用于泛视频业务的层次化编码模型(以空间域为例)

2)将P^o间隔地进行空间域下采样,得到L路子信号以P^l表示,则子路信号空间分辨率为(I, J),其中$I \cdot J = \dfrac{X \cdot Y}{L}$,如遇不能整除的情况,可用超分辨率技术转换$P^o$的分辨率.则视频第$l$路信号的第$k$帧中的任意像素可以记为

$$P^l_{k,i,j},\ 1 \leqslant l \leqslant L,\ 1 \leqslant k \leqslant K,\ 1 \leqslant i \leqslant I,\ 1 \leqslant j \leqslant J. \tag{1}$$

3)将L路子信号进行主成分提取,子路信号主成分提取的方法可以表示为

$$P^{\text{Main}}_{k,i,j} = \frac{1}{L}\sum_{l=1}^{L} P^l_{k,i,j},\ \forall k, i, j, \tag{2}$$

其中P^{Main}表示提取的主路信号.

4)各子路信号利用残差器通过对应元素相减法将主成分去除,此过程可表示为

$$P^{l*}_{k,i,j} = P^l_{k,i,j} - P^{\text{Main}}_{k,i,j},\ \forall k, i, j, l. \tag{3}$$

相应地,视频解码为编码的逆过程,如图3所示.其中,主描述层的编解码可以兼容H.265高效视频编码(high efficient video coding, HEVC)、H.264高级视频编码(advanced video coding, AVC),增强描述层编解码可以兼容SHVC, SVC编解码器,但与之不同的是,增强描述层之间不相互参考,而是仅参考主描述层,如此以提升视频编解码的灵活性.

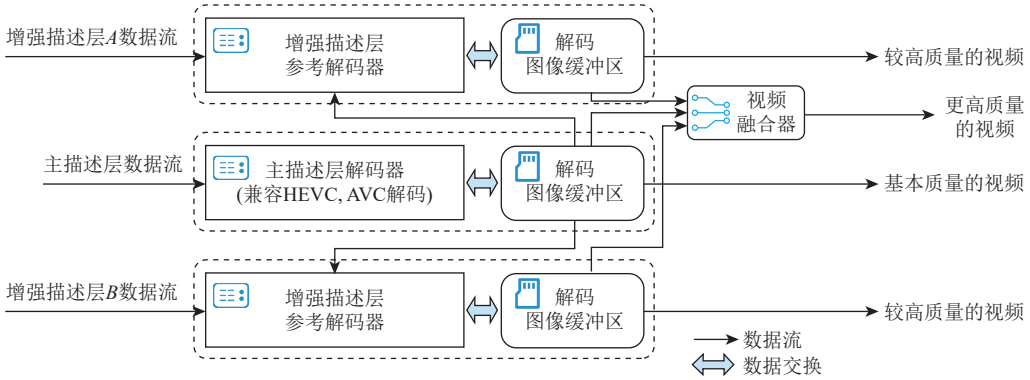

Fig. 3 Hierarchical video decoder block diagram

图3 层次化视频解码器框图

2.2 面向业务的有效吞吐量模型

考虑到算力网络的特性和泛在化视频业务传输约束的差异性,本文提出面向业务的有效吞吐量模型.首先,将一个传输链路记为一个由源m到目的n的信道$(m-n)$,其中源m只标记为可以提供服务的云中心、MEC服务器,或者任何一种网络设备,但不指定其类型.即,在算力网络中,不特定服务的提供者,而关注如何聚合足够的网络资源来提供服务.然后,定义信道质量信息(channel quality information, CQI):RTT_{mn}为从源到目的的往返时延,μ_{mn}为信道$(m-n)$的可用信道容量,γ^c_{mn}为该信道的信道丢包率,为独立、同分布的(i. i. d)过程,可通过链路状态估计得出.本文采用Gilbert信道模型[21]来描述信道丢包率:

$$\gamma^c_{mn} = \frac{\rho^{\text{B,G}}_{mn}}{\rho^{\text{G,B}}_{mn} + \rho^{\text{B,G}}_{mn}},\ \forall m, n, \tag{4}$$

其中,$\rho^{\text{B,G}}_{mn}$为信道由好(G)到坏(B)的状态转移概率,$\rho^{\text{G,B}}_{mn}$反之.

此外,不同于其他类型的业务,视频服务需要考虑数据包的时延约束.即,实时的视频通信业务,如果一个数据包在到达之前,其从属的视频图像组(group of pictures, GoP)已经解码并播放完毕,则该数据包已无效,当以丢包处理.通常,将这种在规定时间内成功到达接收端的数据量定义为有效吞吐量(goodput),并且考虑到不同类型视频业务对延时约束的差异化需求,本文定义这种因未在指定到达时间内到达的数据丢包为超时丢包γ^{o}_{mn},可通过M/G/1排队模型来近似[17,20-21],即:

$$\gamma^{\text{o}}_{mn} = P(T_{mn} > \tau^{\alpha}) = \exp(-\lambda_{mn} \times \tau^{\alpha}), \tag{5}$$

其中,T_{mn}为链路的端到端延时,τ^{α}为不同业务要求的传输截止时间,λ_{mn}为到达率,$\lambda_{mn} = 1/E(T_{mn})$,$E(T_{mn})$为延时的期望值.实际应用中,$E(T_{mn})$可通过历史观测获得,但考虑到该过程较为复杂,本文引入分数函数来近似:

$$E(T^l_{mn}) = \frac{r^l_{mn}}{\omega^l_{mn}} + \frac{RTT_{mn}}{2}, \tag{6}$$

其中,r^l_{mn}为第l层视频信号的码率,ω^l_{mn}为在链路$(m-n)$上为其分配的传输速率.由式(5)和式(6)可得超时丢包率为

$$\gamma^{\text{o}}_{mn} = \exp\left(-\frac{2\tau^{\alpha} \times \omega^l_{mn}}{2r^l_{mn} + \omega^l_{mn} \times RTT_{mn}}\right), \tag{7}$$

进一步，总丢包率为

$$\Psi_{mn}=\gamma_{mn}^{c}+(1-\gamma_{mn}^{c})\times\gamma_{mn}^{o}=\frac{\rho_{mn}^{\mathrm{B,G}}}{\rho_{mn}^{\mathrm{G,B}}+\rho_{mn}^{\mathrm{B,G}}}+\frac{\rho_{mn}^{\mathrm{G,B}}}{\rho_{mn}^{\mathrm{G,B}}+\rho_{mn}^{\mathrm{B,G}}}\times\exp\left(-\frac{2\tau^{\alpha}\times\sum_{l}\omega_{mn}^{l}}{2\times\sum_{l}r_{mn}^{l}+RTT_{mn}\times\sum_{l}\omega_{mn}^{l}}\right),\tag{8}$$

其次，视频在信道$(m-n)$上传输的总有效吞吐量为

$$\Theta_{mn}=\sum_{l}r_{mn}^{l}\times(1-\Psi_{mn}).\tag{9}$$

3 视频传输调度策略

基于本文提出的视频编解码模型和有效吞吐量模型，本节介绍如何根据算力网络中网络资源的分布情况以及用户对视频业务的请求情况对视频的传输进行合理的调度，以提升视频业务的服务质量.

为了保证视频传输的稳定性，在某一链路$(m-n)$上分配的视频流总速率之和不应超过其实际可用容量，即

$$\sum_{l}\omega_{mn}^{l}\leqslant\mu_{mn},\tag{10}$$

其中，ω_{mn}^{l}为在信道$(m-n)$上为第l层视频分配的传输速率，μ_{mn}为信道$(m-n)$的实际可用容量.

$$\begin{cases}\boldsymbol{\mathcal{H}}=\begin{bmatrix}\boldsymbol{\mathcal{H}}^{B}\\ \boldsymbol{\mathcal{H}}^{E}\end{bmatrix},\ \boldsymbol{\mathcal{H}}^{B}=(h_{11},h_{12},\cdots h_{1D}),\ \boldsymbol{\mathcal{H}}^{E}=\begin{pmatrix}h_{21}&\cdots&h_{2D}\\ \vdots&\ddots&\vdots\\ h_{L1}&\cdots&h_{LD}\end{pmatrix}\\ \boldsymbol{\mathcal{H}}=(\ \boldsymbol{H}_1,\quad\boldsymbol{H}_2\quad\cdots\quad\boldsymbol{H}_D\ ),\ \boldsymbol{H}_n=\begin{pmatrix}h_{1n}\\ \vdots\\ h_{Ln}\end{pmatrix},\ \forall n\in\mathcal{D}=\{1,2,\cdots,D\},\end{cases}\tag{12}$$

其中，$\boldsymbol{\mathcal{H}}^{B}$和$\boldsymbol{\mathcal{H}}^{E}$用以表示网络中主描述层和增强描述层的分布，$\boldsymbol{H}_n$则进一步表示$n$中视频层的具体存储状态.

3.1 视频内容按需部署

本文假设从属于同一算力网络中的D个设备以不同分辨率、不同类型(如2D视频、3D视频、全景视频等)同时请求同一个视频内容，并组成集合$\mathcal{D}$. 基于本文提出的视频编码模型，将视频内容的分发分为3个阶段，如图4所示.

第一阶段为视频主描述层的分发阶段，此时所有设备中没有缓存任何视频数据包，所有视频的数据包都是被需要的. 为了保证视频服务的流畅性，所有发起请求的设备将被分配恰当的内容“源”，以广播、多播或点对点形式传输视频的主描述层数据包(B)，此过程结束时，式(13)将被满足.

为方便陈述，首先声明如下参数：记$\boldsymbol{\mathcal{M}}(D\times D)$为算力网络中的可用传输链路矩阵，其中的元素标定了对应链路的可用容量，如μ_{mn}；以0-1矩阵$\boldsymbol{\mathcal{C}}(D\times D)$表示网络中的供需关系，例如若$c_{mn}=1$，则表示网络中的设施$m$可为设施$n$提供所需的视频内容资源；本文假设一个视频被编码为$L$个视频层，并将视频层的速率分配方案用矩阵$\boldsymbol{\Omega}_n(L\times D)$表示，例如如果其中元素$\omega_{mn}^{l}\neq 0$，则表示在信道$(m-n)$上分配了第$l$层的视频数据，且分配速率为$\omega_{mn}^{l}$；对于一个设备(设施)的存储状态，则以0-1矩阵$\boldsymbol{\mathcal{H}}(L\times D)$表示，其中如果$h_{ln}=1$则代表设备(设施$n$)中已经存储了视频第$l$层的数据. 矩阵$\boldsymbol{\mathcal{M}}$, $\boldsymbol{\mathcal{C}}$, $\boldsymbol{\Omega}_n$和$\boldsymbol{\mathcal{H}}$的形式如式(11)所示：

$$\boldsymbol{\mathcal{M}}=\begin{pmatrix}\mu_{11}&\mu_{12}&\cdots&\mu_{1D}\\ \mu_{21}&\ddots&\ddots&\vdots\\ \vdots&\mu_{mn}&\ddots&\vdots\\ \mu_{D1}&\cdots&\cdots&\mu_{DD}\end{pmatrix},\boldsymbol{\mathcal{C}}=\begin{pmatrix}c_{11}&c_{12}&\cdots&c_{1D}\\ c_{21}&c_{22}&\ddots&\vdots\\ \vdots&c_{mn}&\ddots&\vdots\\ c_{D1}&\cdots&\cdots&c_{DD}\end{pmatrix},$$

$$\boldsymbol{\Omega}_n=\begin{pmatrix}\omega_{1n}^{1}&\omega_{2n}^{1}&\cdots&\omega_{Dn}^{1}\\ \omega_{1n}^{2}&\omega_{2n}^{2}&\ddots&\vdots\\ \vdots&\omega_{mn}^{l}&\ddots&\vdots\\ \omega_{1n}^{L}&\cdots&\cdots&\omega_{Dn}^{L}\end{pmatrix},\boldsymbol{\mathcal{H}}=\begin{pmatrix}h_{11}&h_{12}&\cdots&h_{1D}\\ h_{21}&h_{22}&\ddots&\vdots\\ \vdots&h_{ln}&\ddots&\vdots\\ h_{L1}&\cdots&\cdots&h_{LD}\end{pmatrix}.\tag{11}$$

进一步地，鉴于视频的主描述层与增强描述层具有不同的性质，$\boldsymbol{\mathcal{H}}$矩阵可记为如式(12)所示的2种形式：

$$\left\|\boldsymbol{\mathcal{H}}^{B}\right\|_{1}=|\mathcal{D}|.\tag{13}$$

第二阶段为视频的增强描述层的分发阶段，此

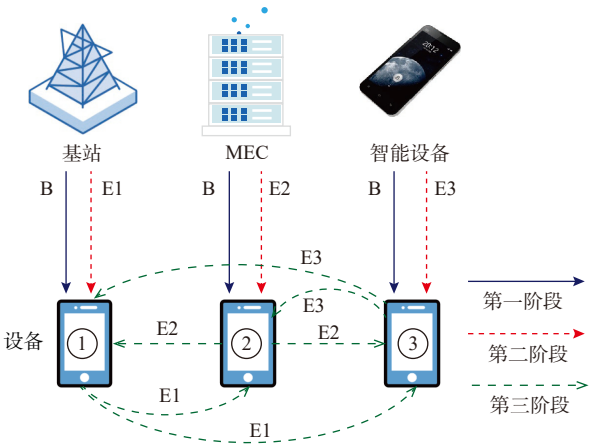

Fig. 4 Video content distribution model

图4 视频内容分发模型

时所有设备中均没有缓存增强层数据，任意增强层数据均可以为设备提升视频的播放质量，但为了提升算力网络中“碎片化”网络资源的利用率，还需对增强层数据包的部署进行统一规划. 因此该阶段的主要目的有 2 个：一是为了尽力提升视频业务的服务质量，为服务的请求者分配恰当的“源”来提供其所需的视频增强描述层数据包(如 E1,E2,E3)，二是为了提升算力网络中分布的碎片化网络资源，当前阶段对视频内容的部署应该便于下一阶段的视频内容分享. 例如，在图 4 中，当设备③请求一个增强描述层数据包时，无论算力中心将哪一个增强层数据包分配给设备③都可以提升其视频业务质量. 但考虑到在设备③的连接范围内，设备①和设备②缺失的数据包是 E3，因此将 E3 分配给设备③最能促进设备③的资源分享. 记$f(n,l)$为视频内容分发增益度函数，其代表了给设备 n 分配第 l 层数据能够产生的资源分享增益. 例如 $f(3,\text{E3})=2$. 则该过程可以用最优化函数来表示：

$$\max \sum_{n\in\mathcal{D},\, l\in\Delta_n} f(n,l), \tag{14}$$

其中，Δ_n为设备 n 缺失的视频层集合，可通过查询$\boldsymbol{H}_n$获得. 式(14)所示的最优化过程可以通过算法 1 来实现.

算法 1. 最优视频层分配算法.

输入：$\mathcal{H}$, $\mathcal{D}$, $\Delta_n, n\in\mathcal{D}$, 视频所有 L 层数据；

输出：$\mathcal{S}^{\mathrm{E}}$.

① $\mathcal{S}^{\mathrm{E}}=[0]$;

② while $\mathcal{D}\neq\varnothing$ do

③ 随机选择一个$n\in\mathcal{D}$;

④ for all $l\in\Delta_n$ do

⑤ 初始化 $f(n,l)=0$;

⑥ 将在 n 的一跳范围之内的设备记为 n';

⑦ for all n' do /* 计算 $f(n,l)$ */

⑧ if $l\in\Delta_{n'}$

⑨ $f(n,l)=f(n,l)+1$;

⑩ end if

⑪ end for

⑫ 输出 $f(n,l)$;

⑬ end for

⑭ 在所有 $f(n,l)$中查询l^*，使得 $f(n,l^*)$的值在所有 $f(n,l), l\in\Delta_n$中最大；

⑮ 选择l^*分配给设备 n;

⑯ $s_{l^*n}\leftarrow 1$;

/* 将$\mathcal{S}^{\mathrm{E}}$中对应元素更新为 1，以记录分配策略 */

⑰ 将l^*发送给 n，更新$\boldsymbol{H}_n$;

⑱ 返回 ACK, CQI, $\boldsymbol{H}_n$给计算中心，更新$\mathcal{H}$;

⑲ $\mathcal{D}\leftarrow\{n\}/\mathcal{D}$;

/* 将 n 从集合$\mathcal{D}$中移除 */

⑳ end while

算法 1 的输出$\mathcal{S}^{\mathrm{E}}$即为视频层的最优分配方案.$\mathcal{S}^{\mathrm{E}}(L\times D)$为 0-1 矩阵，其形式为：

$$\mathcal{S}^{\mathrm{E}}=\begin{pmatrix} s_{11} & 0 & \cdots & s_{1D} \\ 1 & 0 & \ddots & \vdots \\ \vdots & s_{ln} & 1 & \vdots \\ s_{1L} & \cdots & \cdots & s_{LD} \end{pmatrix}. \tag{15}$$

第三阶段为协作式的视频内容分享阶段，此时不同设备中可能存储了不同的视频增强描述层数据包，不同设备之间彼此分享，在为其它设备提供已有数据包的同时，从其它设备中获取缺失的数据包，共同提升视频的播放质量. 该阶段可视为将存储了一定内容的视频业务请求者加入了提供者的集合. 只要某一些设备被算法 1 或算法 2 选择，则这些设备将同时扮演请求者和服务提供者的身份，分享自身的网络资源.

3.2 基于有效吞吐量的链路选择

经过视频内容的分发后，视频以一定的形式分布于网络中，例如一些设备、设施中存储了视频的若干层或者若干视频片段，接下来便是如何有效地利用这些分散的视频内容，在提升视频业务质量的同时，提高碎片化网络资源的利用率. 由于本文提出的视频编码模型中，增强描述层是同等重要的，因此设备 n 并不指定增强描述层编号，而由计算中心根据全局调度来决定具体发送给设备 n 哪一个增强描述层. 当一个设备 n 请求一个视频增强描述层时，首先需要确定的是能为设备 n 提供增强描述层的设备(设施)的集合. 记$\boldsymbol{H}_m$为 m 中存储的视频层集合，即存储状态向量$\boldsymbol{H}_m$中所有 1 元素对应位置的坐标. 记$\mathcal{P}_n'$为存储了 n 所需视频层的设备(设施)的集合，即$\mathcal{P}_n'$为那些使得$\boldsymbol{H}_m\cap\Delta_n\neq\varnothing$的设备(设施)的集合. 除此之外，考虑到网络中单个设备(设施)的有限能力，在为设备 n 提供服务之前，还需确定该设备是否“空闲”，例如是否具有足够的可用带宽、是否电量充足等. 因此，实际的内容提供者集合定义为$\mathcal{P}_n=\mathcal{P}_n'\cap\Xi$，其中$\Xi$为“空闲”设备(设施)的集合.

接下来便是如何选择一个合适的提供者并发送合适的视频层给设备 n. 考虑到不同类型的视频业务的不同性质(例如对延时约束的不同要求)，本文基

于提出的有效吞吐量模型提出业务需求约束下的链路选择策略，如图 5 所示. 得益于算力网络的支撑，本文提出的链路选择算法可以基于对全局的观测实现最优的视频传输调度. 首先，当一个设备请求一个视频层时，它将自己缺失的视频层集合以及具体的业务约束（τ^{α}）发送给计算中心. 其次，计算中心综合全局的$\mathcal{M}$, $\mathcal{C}$、$\boldsymbol{\Omega}_n$, $\mathcal{H}$、Ξ以及个体的Δ_n, τ^{α}生成最优的调度策略，该调度策略主要有 2 个功能：1）根据网络中视频内容的分布情况，决定发送给 n 的视频层编号 l*，使得视频内容的分布更利于碎片化网络资源的利用. 2）根据 n 的视频业务约束和有效吞吐量模型，为 n 分配合适的传输链路，保证视频业务质量. 最后，依照最优化的调度决策，由计算中心指导网络中的设备或设施开展视频数据传输，一方面，将复杂的计算和决策工作分配给计算中心执行，实现全局最优；另一方面，数据的传输以分布式的方式进行，便于网络中碎片化的网络资源灵活利用. 本文提出的基于有效吞吐量的链路选择算法伪代码如算法 2 所示.

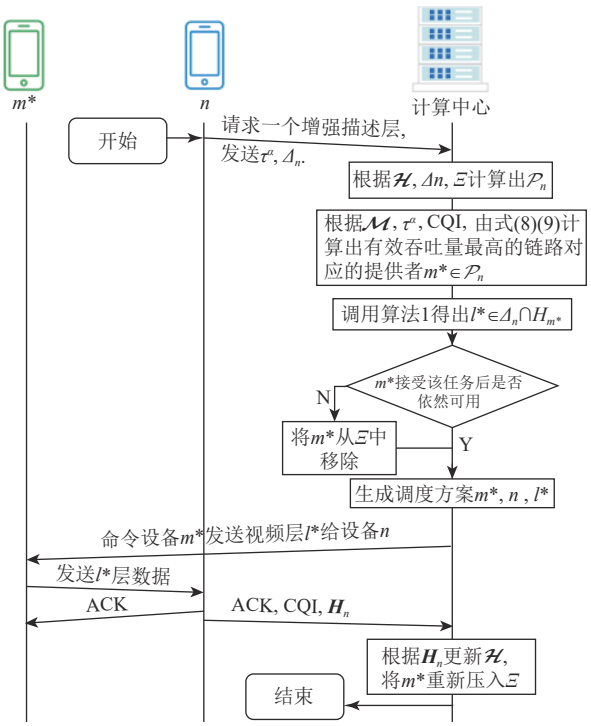

Fig. 5 Link selection strategy under service requirement constraints

图 5 业务需求约束下的链路选择策略

算法 2. 基于有效吞吐量的链路选择算法.

输入：$\mathcal{M}$, $\boldsymbol{\Omega}_n$, $\mathcal{H}$, Ξ;

输出：$\mathcal{S}$.

① while 设备 n 发起一个请求，发送Δ_n, τ^{α} do

② t=0;

③ while $t < \tau^{\alpha}$或者$\|\boldsymbol{H}_n\|_1 = L$

④ 根据$\mathcal{H}$、Ξ、Δ_n计算$\mathcal{P}_n = \mathcal{P}_n' \cap \Xi$;

/* 即通过查询供需关系、空闲状态确定 n 的提供者集合 */

⑤ for all $m \in \mathcal{P}_n$ do

⑥ 根据$\mathcal{M}$、$\boldsymbol{\Omega}_n$, 由式（8）和式（9）计算Θ_{mn};

/* 即通过信道状态预测有效吞吐量 */

⑦ end for

⑧ 求出最大Θ_{mn}对应的 m*;

⑨ 调用算法 1 求出 l*，生成调度方案 $\mathcal{S} = (m*, n, l*)$;

/* 由 m*为 n 提供 l* */

⑩ 令 m*发送 l*给 n，将 l*从Δ_n中移除，更新$\boldsymbol{H}_n$、$\mathcal{H}$、Ξ、t;

⑪ end while

⑫ end while

4 仿真和数值分析

为了衡量本文提出方案的性能，本节从核心网负载、有效吞吐量、碎片化网络资源利用率等 3 个方面将本文方案与 DASH 方案[9]、基于 MEC 的有效吞吐量方案（MEC-enabled goodput-aware, MEGA）[17]、基于本文编码模型的随机调度方案以及基于 SVC 编码的全局调度方案等 4 种方案进行了仿真比较.仿真环境为：操作系统（Windows 10 Enterprise 64 b），CPU（英特尔 Core i7-8700 @ 3.20 GHz 6 核），内存（8 GB DDR4 2666MHz），GPU（Nvidia Quadro K4000 3 GB），仿真软件（MATLAB R2020a-academic use）. 如无特殊说明，本节中的仿真参数默认如表 1 所示.

Table 1 Simulation Parameter Values Setting

表 1 仿真参数值设置

参数	数值
视频层数 L	10
请求设备数 D	200
业务类型数	10
RTT/ms	50
视频码率/Mbps	1
信道带宽/MHz	1
信道丢包率/%	5

如图 6 所示，网络中的设备从不同时间开始，随机地发起视频服务请求. 请求的类型数为 10 种，即不同的质量（层数 L）和不同的延时约束（τ^{α}）. 由于 DASH 方案为集中式的视频服务模式，核心网承载了所有

的流量负担，因此随着时间的变化，网络中请求视频的设备数目逐渐增多，核心网的流量负载基本上呈线性提升. 其余的 4 种方案均可以支持分布式的视频传输模式，因此均可以有效降低核心网的负载. 但在起始阶段，由于网络中还不存在可用的视频数据包，因此核心网仍然承载了全部流量负担. 随着时间的推移，越来越多的网络设备(设施)中存储了可用的视频内容，因此利用这些设备(设施)中的碎片化网络资源来分享自身存储的视频内容，便有效降低了核心网的流量负担.

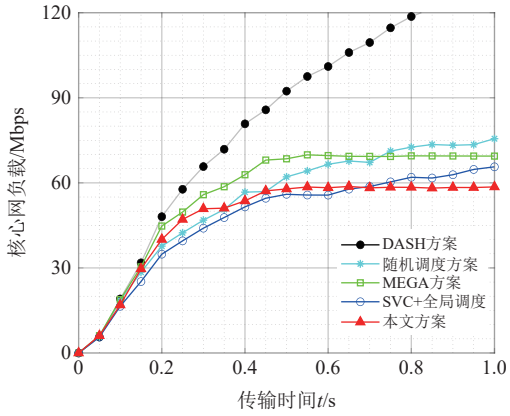

Fig. 6 Changes of core network load with time

图 6 核心网负载随时间的变化

值得注意的是，在传输的起始阶段，由于网络中分布的视频内容有限，核心网需要分发更多的视频数据包到网络中. 而由于本文提出的视频编码模型为了提升码流的灵活性，相对 SVC 而言牺牲了部分编码效率(为进行增强描述层之间的参考编码)，因此在这个过程中，核心网负载相对较高. 随着网络中分布的视频数据包越多，本文提出的方案的优势也越明显. 本文提出的方案优于其它 3 种分布式方案的主要原因：1)全局的最优化调度利于碎片化资源的有效利用，因为设备所需的资源更容易在网络设备而非核心网设施中获得. 因此本文方案优于随机调度方案和 MEGA 方案. 2)本文提出的视频编码模型更具灵活性，当设备需要提升视频质量时，是请求了一个增强层的集合而非特定的某一层，使得其更容易在网络设备中获得. 因此本文方案可以优于基于 SVC 的全局调度方案.

图 7 展示了本文方案与对比方案在不同码率下，核心网稳定负载(传输进行 10s 以后)的对比情况. 5 种方案核心网负载均基本上与视频码率呈线性正相关关系，且由于本文方案更容易进行分布式视频分享，因此在卸载核心网负载方面更有效.

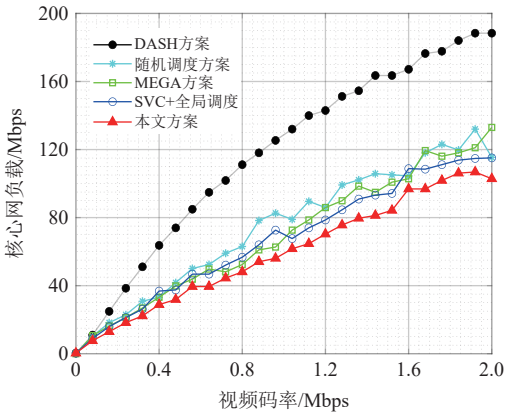

Fig. 7 Change of core network load with video bit rate

图 7 核心网负载随视频码率的变化

图 8 展示了有效吞吐量指标随时间的变化情况对比. 由于本文方案与先前工作中(MEGA 方案)使用了相似的有效吞吐量模型，因此二者在有效吞吐量方面差别不大. 基于 SVC 的全局调度方案由于仅从网络设备中获取视频内容的概率稍低，因此在有效吞吐量方面稍逊于本文方案和 MEGA 方案. 集中式的 DASH 方案由于只支持核心网络设施到设备的链接，因此有效吞吐量指标表现较差，但仍优于随机的调度方案. 这是因为随机的调度方案容易使得设备所需内容被分配到不稳定的信道，使得有效吞吐量指标波动剧烈.

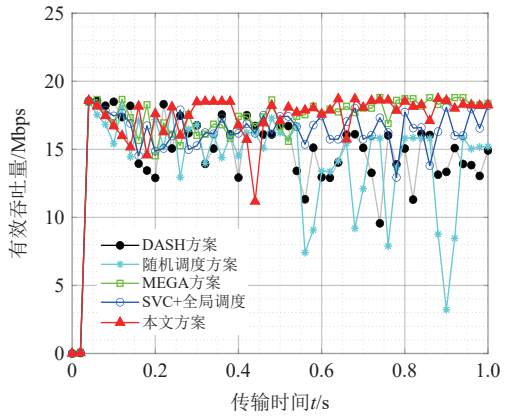

Fig. 8 Changes of good-put with time

图 8 有效吞吐量随时间的变化

图 9 展示了有效吞吐量指标在不同视频码率下的对比，变化的原因与图 8 类似. 受用户请求的随机性、无线信道的动态性的影响，视频内容的部署仍具有随机特征，因此有效吞吐量指标不可避免出现波动特征，但总体而言，本文算法可以取得最好的性能.

图 10 展示了本文算法在提升碎片化网络资源利用率方面的性能，对比仿真在 4 种分布式的视频内

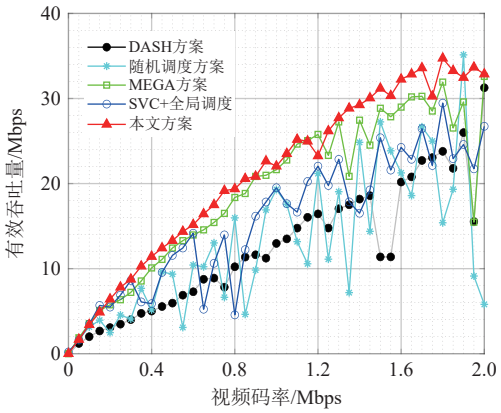

Fig. 9 Changes of good-put with video bit rate

图 9 有效吞吐量随视频码率的变化

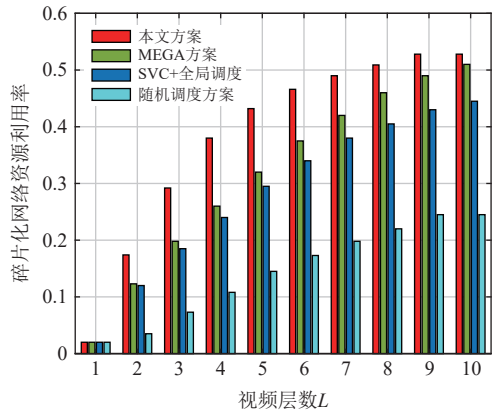

Fig. 10 Comparison of utilization rates of fragmented network resources

图 10 碎片化网络资源利用率情况对比

容分享方案. 该仿真中，本文主要统计了网络设备(如智能终端)中碎片化网络资源的利用情况. 只要某一设备分享了其中的视频内容，则视为该设备的碎片化资源得到了有效利用，核心网设施(如基站、MEC 服务器)不计入其内. 由仿真结果可以看出，在视频层数为 1(即不分层)时，4 种方案对碎片化网络资源的利用率一致且较低，因为视频内容的分布和有限的终端设备能力使得设备间的协作共享较难进行. 随着视频层数的提升，视频内容较容易从邻居设备中获得，因此碎片化的网络资源利用率得以有效提升. 得益于本文提出的灵活的视频编码模型和全局最优的视频调度，本文方案使得设备间的协作容易开展，对比其它 3 种方案，本文方案能使碎片化网络资源利用率最高.

5 结 论

本文提出了一种算力网络支撑下的泛在化视频业务调度方案，以缓解爆炸式增长的泛在化视频业务数据量对核心网承载能力的挑战. 具体而言，本文首先通过提出一种适用于泛视频业务的层次化视频编码模型，可以在适应无线网络动态性、用户和业务需求多样性的同时，使得视频内容便于灵活地传输；其次通过提出一种基于全局观测和有效吞吐量模型的视频调度策略，本文方案可以有效提升碎片化网络资源的利用率，从而卸载核心网负载. 仿真结果表明了本文方案的有效性.

未来的工作包括 2 个方面：1)发掘用户行为习惯特征，利用人工智能和大数据方法实现对视频内容的智能部署；2)发掘泛在化视频应用中用户体验与网络资源配置的潜在联系，实现网络资源的自适应配置.

作者贡献声明：张旭光负责完成实验并撰写论文；陈鸣锴提出指导意见并修改论文；魏昕提出了算法思路和实验方案.

参 考 文 献

[1] Dong Yu, Song Li, Xie Rong, et al. Ultra-low latency, stable, and scalable video transmission for free-viewpoint video services[J]. IEEE Transactions on Broadcasting, 2022, 68(3): 636−650

[2] Minopoulos G, Psannis K E, Kokkonis G, et al. QoE assessment of video codecs for video streaming over 5G networks[C] //Proc of the 2020 3rd World Symp on Communication Engineering (WSCE). Piscataway, NJ: IEEE, 2020: 34−38

[3] Yaqoob A, Bi Ting, Muntean G M. A survey on adaptive 360 video streaming: Solutions, challenges and opportunities[J]. IEEE Communications Surveys & Tutorials, 2020, 22(4): 2801−2838

[4] Hofbauer M, Kuhn C B, Khlifi M, et al. Traffic-aware multi-view video stream adaptation for teleoperated driving[C] //Proc of the 2022 IEEE 95th Vehicular Technology Conf: Piscataway, NJ: IEEE, 2022: 1−7

[5] Shutsko A. User-generated short video content in social media: A case study of TikTok[C] //Proc of the Int Conf on Human-Computer Interaction. Switzerland: Springer, Cham, 2020: 108−125

[6] Lei Bo, Liu Zengyi, Wang Xuliang, et al. Computing network: A new multi-access edge computing[J]. Telecommunications Science, 2019, 35(9): 44−51 (in Chinese)
(雷波, 刘增义, 王旭亮, 等. 基于云、网、边融合的边缘计算新方案：算力网络[J]. 电信科学, 2019, 35(9): 44−51)

[7] Tang Xiongyan, Cao Chang, Li Jianfei, et al. Cutting edge report of computing power network[R]. Beijing: China Communication Society, 2020
(唐雄燕, 曹畅, 李建飞, 等. 算力网络前沿报告[R]. 北京: 中国通信学会, 2020)

[8] Seufert M, Egger S, Slanina M, et al. A survey on quality of experience of HTTP adaptive streaming[J]. IEEE Communications Surveys & Tutorials, 2017, 17(1): 469−492

[9] Deng Zhenjie, Liu Yanwei, Liu Jinxia, et al. Cross-layer DASH-based multipath video streaming over LTE and 802.11ac networks[J]. Multimedia Tools and Applications, 2021, 80(10): 1573−7721

[10] Chen Youjia, Cai Yuekai, Zheng Haifeng, et al. Cooperative caching for scalable video coding using value-decomposed dimensional networks[J]. China Communications, 2022, 19(9): 146−161

[11] Zhan Cheng, Wen Zhe. Content cache placement for scalable video in heterogeneous wireless network[J]. IEEE Communications Letters, 2017, 21(12): 2714−2717

[12] Abiri M, Mehrjoo M, Rezaei M. Scalable video traffic offloading for streaming services in 5G HetNets[J]. Multimedia Tools and Applications, 2022, 81(9): 12325−12347

[13] Li Qi, Nayak A, Wang Xiaoxiang, et al. A collaborative caching-transmission method for heterogeneous video services in cache-enabled terahertz heterogeneous networks[J]. IEEE Transactions on Vehicular Technology, 2022, 71(3): 3187−3200

[14] Chatterjee S, De S. QoE-aware cross-layer adaptation for D2D video communication in cooperative cognitive radio networks[J]. IEEE Systems Journal, 2022, 16(2): 2078−2089

[15] Liu Jianlong, Wen Jiaye, Lin Lixia, et al. Double agents-DQL based D2D computing-offloading for SHVC[J]. Peer-to-Peer Networking and Applications, 2022, 15(1): 56−76

[16] Zhang Xuguang, Wei Xin, Zhou Liang, et al. Social-content-aware scalable video streaming in Internet of video things[J]. IEEE Internet of Things Journal, 2022, 9(1): 830−843

[17] Zhang Xuguang, Lin Huangda, Chen Mingkai, et al. MEC-enabled video streaming in device-to-device networks[J]. IET Communications, 2020, 14(15): 2453−2461

[18] Wu Liang, Zhang Wenyi. Caching-based scalable video transmission over cellular networks[J]. IEEE Communications Letters, 2016, 20(6): 1156−1159

[19] Zhang Xuewei, Ren Yuan, Lv Tiejun, et al. Caching scalable videos in the edge of wireless cellular networks[J]. IEEE Network, 2022: 1−9

[20] Zhang Xuguang. Research on video communication technologies oriented to D2D networks[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2021
(张旭光. 面向D2D网络的视频通信技术研究[D]. 南京: 南京邮电大学, 2021)

[21] Wu Jiyan, Yuen Chau, Cheng Bo, et al. Goodput-aware load distribution for real-time traffic over multipath networks[J]. IEEE Transactions on Parallel and Distributed Systems, 2015, 26(8): 2286−2299

Zhang Xuguang, born in 1988. PhD. His main research interest includes multimedia communications and computing.

张旭光, 1988 年生. 博士. 主要研究方向为多媒体通信及计算.

Chen Mingkai, born in 1989. PhD. His main research interests include multimedia communications and computing, resource allocation, signal processing in wireless networks.

陈鸣锴, 1989 年生. 博士. 主要研究方向为多媒体通信与计算、资源分配、无线网络中的信号处理.

Wei Xin, born in 1983. Professor, PhD supervisor. His main research interests include multimedia communications and computing, educational technology.

魏　昕, 1983 年生. 教授, 博士生导师. 主要研究方向为多媒体通信与计算、教育科技.