计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330023
60(4): 828−838, 2023

视频云网平台中智能算法版权管理方法

张欢欢[1] 安聪凯[1] 赵朗程[1] 周安福[1] 马华东[1] 袁 艺[2] 曹 宁[2]
[1](北京邮电大学计算机学院(国家示范性软件学院) 北京 100876)
[2](中电信数智科技有限公司 北京 100035)
(zhanghuanhuan@bupt.edu.cn)

Algorithmic Intelligence Right Management Method in Video Cloud-Network Platform

Zhang Huanhuan[1], An Congkai[1], Zhao Langcheng[1], Zhou Anfu[1], Ma Huadong[1], Yuan Yi[2], and Cao Ning[2]
[1](*School of Computer Science* (*National Pilot Software Engineering School*), *Beijing University of Posts and Telecommunications*, *Beijing* 100876)
[2](*China Telecom Digital Intelligence Technology Co., Ltd. Beijing* 100035)

Abstract Video cloud-network platform contains a huge amount of intelligent algorithms, and it is an important scientific problem to efficiently manage video cloud-network platform so as to support the rapid deployment and update of application services. However, the traditional intelligent algorithms are forcibly bounded to cloud resource, and there is no unified invocation mechanism for intelligent algorithms among different service providers, which is difficult for fast integration and effective utilization. In order to solve this problem, we propose the "service-algorithm-resource" dynamic interconnection service system, which can effectively solve the contradiction among rapid iteration of algorithm, dynamic application demand and fixed management of intelligent algorithms. In the process of dynamic interconnection services, the traditional and fixed content-oriented buyout digital rights management can no longer provide efficient services for fine-grained rights management. To this end, we propose an algorithmic intelligence right management (AIRM) system, and build "sharing-mode" intelligent algorithm right management method on the video cloud-network platform through right resource servitization method and liquidity arithmetic network structure. The actual deployment results in the China telecom video analysis platform authorization management module show that the designed method can increase the parallel algorithm service capacity by 19.9 times, and decrease the right response time by 18.36%.

Key words video cloud-network platform; right management; algorithmic intelligence; service scheduling; computing power network

摘 要 视频云网平台中涵盖了大量智能算法,如何对其进行高效管理,从而支持应用服务的快速部署与更新是一个重要的科学问题.然而,传统的智能算法与云端资源具有绑定规则,不同应用服务商之间的智能算法缺乏统一的调用机制,导致它们无法快速整合和有效利用.为了解决此难题,建立"服务—算法—

收稿日期: 2023-01-08; **修回日期**: 2023-02-15
基金项目: 国家自然科学基金创新研究群体项目 (61921003); 中国电信视觉智联平台能力引入项目; 博士后创新人才支持计划项目(BX20220046)
This work was supported by the Innovation Research Group Project of the National Natural Science Foundation of China (61921003), the Platform Capability Project of China Telecom Intelligent Visual Internet of Things; and the China National Postdoctoral Program for Innovative Talents (BX20220046).
通信作者: 马华东(mhd@bupt.edu.cn)

资源”动态互联服务体系，有效解决算法快速迭代、应用需求时变与智能算法版权固化管理的矛盾. 在动态互联服务过程中，传统的、面向固定内容的买断式数字版权管理已经无法为细粒度权限管理提供高效服务. 为此，提出智能算法版权管理系统（algorithmic intelligence right management, AIRM），通过设计版权资源服务化方法与流动性算力网络结构，构建视频云网平台中“共享式”智能算法版权管理方法. 在中国电信视频分析平台授权管理模块中的实际部署结果表明，所设计方法可以将算法并发服务能力提高19.9倍，将算法版权响应时间降低18.36%.

关键词 视频云网平台；版权管理；智能算法；服务调度；算力网络

中图法分类号 TP399

近年来，视频物联、视频会议等新型视频网业务快速发展，思科发布的《网络可视化报告》中指出，2022年的视频流量超过了全球互联网下行流量的82%，视频网业务成为了物联网、移动互联网最主要的业务类型. 视频网业务通常需要云端资源与网络资源共同协作来提升视频用户体验质量，视频云网融合对视频用户体验质量保障起到了至关重要的作用. 然而，视频网具有用户设备广泛互联、连接量大、交互性强等特点，对云网平台中的算力资源提出了非常高的管理与计算需求. 特别地，在视频云网平台中，涵盖了大量智能算法，例如实时视频监控中的智能车辆识别算法、智能人脸识别算法等. 面对大量的、异构的智能算法，如何对其进行有效管理，从而支持应用服务的快速部署是一个重要的科学问题. 然而，传统的智能算法与云端资源具有绑定规则，不同服务商之间的智能算法缺乏统一的调用机制，导致它们无法快速整合和有效使用. 对于用户而言，智能算法的部署繁琐且过于专业化，用户很难自主完成智能算法与云端资源的装载以及授权过程；对于算法提供商而言，过于专业化的实现过程导致其算法版权很难得到广泛推广和使用，导致算法的商业价值大大降低.

本文提出“算法即服务”理念，建立了新型的“服务—算法—资源”动态互联服务体系，有望解决视频云网平台中算法快速迭代、应用需求动态时变与算法版权固化管理的矛盾. 进一步地，本文的核心目标是将多家智能算法提供商共同组成的算法生态和快速发展的云节点计算生态“软硬”结合，以更高的效率为终端用户提供广泛的“云—边—端”服务能力.

此外，我们发现支持动态服务互联的关键问题在于快速迭代的人工智能算法需要按量计价的细粒度权限管理，然而，传统的、面向固定内容的“买断式”数字版权管理系统已经无法提供有效支持，导致了算法生态与云生态的割裂. 为此，我们提出智能算法版权管理（algorithmic intelligence right management, AIRM）系统. 结合算法版权管理中细粒度、低时延需求等特性，AIRM针对服务层算法版权管理和资源层计算资源管理，设计了版权资源服务化和流动性算力网络2个主要方法，从而支持“共享式”版权管理模式，具体阐述有2点：

1）版权资源服务化. 设计版权资源按需时域划分方法，将有限的版权资源在软件级别扩张为数倍乃至数十倍的细粒度可用服务单元，提高对智能算法版权资源的利用率. 具体来讲，对于每个算法版权，AIRM系统根据用户鉴权频率确定其时域服务单元划分个数，每个算法版权的授权过程由AIRM系统统一调度. 继而，AIRM系统为每个服务单元定义二级服务密钥，它用于服务单元所属用户获取算法版权的时域授权，从而获得时域连续的算法服务，进而提高智能算法服务效率.

2）流动性算力网络设计. 打破单个节点算力资源与版权资源的僵化匹配模型，将广泛的节点计算资源抽象为算力网络；与版权服务单元相协调，提高系统的整体服务能力. 它包括2个主要步骤：首先，根据上层服务需求，设计经验驱动的算力单元划分方法；其次，将云端物理资源切分为细粒度的算力单元作为资源调度单位，多个算力单元可以再自由组合和交换，为多个服务单元提供流动性计算能力，从而实现更高的并行计算效率.

我们将智能算法版权服务与算力网络环境部署在中国电信视频分析平台授权管理模块，系统服务部署在云平台虚拟化后的虚拟机容器中，实验平台由权限认证服务集群、算法存储服务集群以及邦联平台计算资源池3个组成部分. 实验结果表明，相比于传统的版权管理方案，AIRM系统可以在相同算力资源条件下，将并发服务能力提高19.9倍. 此外，在相同服务量的情况下，AIRM系统可将版权响应时间降低18.36%. 针对2种智能算法组合服务的应用场景，AIRM可以将组合服务量提升8倍，证明了其在多种类算法、多任务处理下的有效性能提升.

1 相关工作介绍

新型视频网业务种类众多，例如城市安防监控中进行人员识别、道路监控中进行车辆识别与追踪[1]，这些应用的部署离不开视频云网平台中智能算法与算力网络的支撑[2]. 本节将分别从算法版权管理与算力网络2个方面介绍相关工作.

1)算法版权管理. 数字版权管理(DRM)是一种保护高价值数字资产并控制其分发和使用的系统，文献[3]中介绍了传统DRM的基本框架，包括内容提供者、内容分发者、交易中介和消费者4种角色. 消费者从内容分发者和交易中介处分别获取加密内容和秘钥以实现内容读取，内容通过非对称加密以进行保护. 在云服务场景下，DRM的私钥仅供一个消费者实例使用，这一设计虽然保护了版权但也导致了算法版权无法在容器间高效流动的问题[4-5]. 近年来，随着人工智能的发展，研究者开始关注于算法模型的加密方法[6-10]. 文献[9]通过在神经网络的标准化层中加入验证机制，以保证算法仅在拥有权限的情况下可用，另一些工作[7,10]通过配置神经网络权重以构建数字水印从而达到版权管理的目的. 文献[3–10]工作提升了DRM智能算法加密、解密效率，而未关注于算法服务场景下DRM的系统架构不足，因此无法解决视频应用中的智能算法种类多、版本迭代快导致的DRM执行效率低等问题. 而本文提出的AIRM系统将在实现算法模型加密的基础上，优化智能算法版权管理架构的执行效率.

2)算力网络. 算力网络为网络视频应用提供了人工智能算法执行所需的算力资源以及网络传输资源. 文献[11]提出将算力与网络资源以机器学习平台的形式提供服务，其中算法、算力均由云计算运营商提供，并将这种方式命名为“机器学习即服务”(MLaaS). 后续工作进一步关注了MLaaS模式下集群的资源负载时间、空间特征以及资源效率优化方法[12-16]，例如通过平台内GPU共享与任务作业时长预测可以提升机器学习算法执行过程中的效率[13]. 然而文献[11–16]工作更关注算法的执行效率而并未考虑算法版权的高效授权过程. 尽管版权保护已被认为是算力网络中亟待解决的问题[17]，除MLaaS相关工作外，算力网络中数字版权的认证流程与权限管理方案已有相关研究[18-22]，然而这些研究中版权所保护的内容如音视频[20]、数据库[22]等往往仅涉及到云环境中的储存资源相关的版权调用. 而本文的AIRM系统将提供算力网络调度与智能算法的版权管理的深度耦合架构，从而提升算力网络的服务能力.

2 设计架构

本文提出了新型的动态互联服务体系，其核心目标是将多家智能算法提供商组成的算法生态和当前飞速发展的云节点计算生态“软硬”结合，以更高的效率为端用户提供广泛的“云—边—端”服务能力. 进一步来说，动态互联服务体系从多家智能算法提供商购入一定数量的多种智能算法版权，备份在本系统内部的算法版权存储器中. 继而，将算法版权重新转换为细粒度的服务单元，为其设计二级秘钥来授权于更多的应用用户. 同时，我们将广泛的云节点计算资源构建为高效的流动性算力网络，由服务单元调用其算力单元完成用户的服务请求，最终实现“算法即服务”的应用体系.

以视频物联中经典应用——城市人员安防——为例，介绍其版权服务系统部署方式与成本开销. 城市中部署了大量的摄像头设备来监控道路安全情况，如果为每一个摄像头都配备算力资源和智能算法版权来实现智能监控，则将面临巨大的部署成本. 例如，2021年中国摄像头已增至5.6亿台，而单个摄像头的算法版权费用和支持图像计算的硬件成本一般在千元级别. 所以，我们希望将众多的端设备通过互联服务体系在云端完成应用服务，此种部署方式将极大地降低基建部署成本和降低部署难度，有潜力提高智能算法服务的自适应性和广泛性.

为实现这一服务体系，本文提出了支持动态互联的系统架构. 如图1所示，该架构实现了算法生态(由多家算法提供商共同构成)和云计算生态(由海量计算资源构成)的动态互联，进而提供更高效和便捷的智能算法服务. 动态互联系统架构被设计为3层架构：应用层、服务层、资源层. 具体来讲，应用层提供与用户直接交互的各种智能算法，它接入和整合了各种算法服务API，向用户提供可操作的用户界面；服务层封装了多样、异构的音视频智能应用算法版权，将多种应用的多个智能算法版权在该层实现集中式管理和细粒度调用；资源层实现了由海量云端节点计算资源互联构成的算力网.

3 算法设计

本文针对服务层算法版权管理和资源层计算资

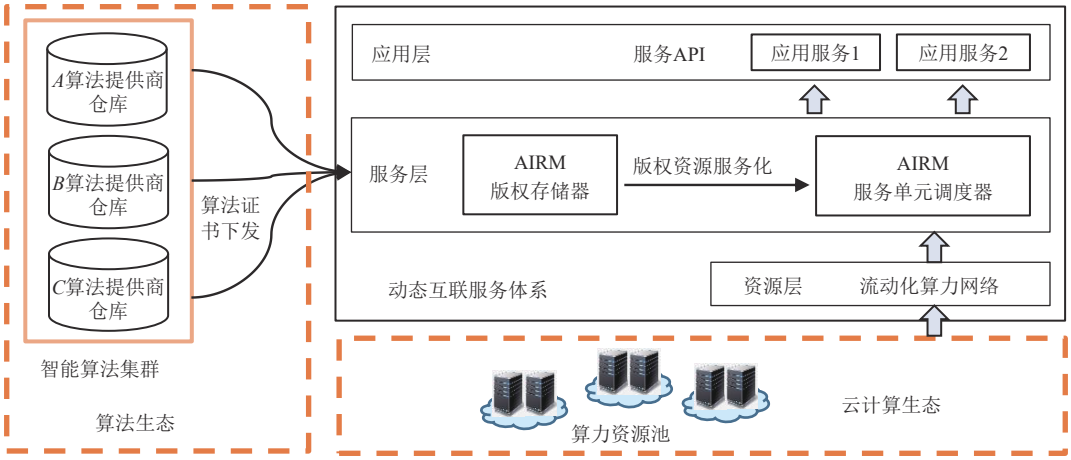

Fig. 1 Dynamic interconnection service system of "Algorithm-Service-Resource"

图 1 “算法-服务-资源”动态互联服务体系

源管理设计了 2 个主要方法：

1)版权资源服务化. 将有限的版权资源在软件级别扩张为数倍甚至数十倍的细粒度可用服务单元，提高对智能算法版权资源的利用率.

2)流动性算力网络. 打破单个节点算力资源与版权资源的僵化匹配模型，将广泛的节点计算资源抽象为算力网络，与版权服务单元相协调，提高系统的整体服务能力.

版权资源服务化与流动性算力网络分别在本服务体系的服务层和资源层实现，后者对前者是透明的，服务单元只需要向下层递交资源请求即可完成调用，而无需关注底部资源层如何实现算力调度，这样的分层设计可以提高服务层的算法服务可扩展性. 接下来将具体阐述 2 个方法的设计原理.

3.1 版权资源服务化

首先，介绍传统“买断式”数字版权管理方法(DRM). 用户从版权提供商购买智能算法版权，并自己组建计算资源进行使用. 这种方式导致算法版权和计算资源高度耦合，只能被单个用户使用，并且要求用户有较高的专业技能来完成部署，严重降低了算法版权的利用率和可用性. 我们发现，用户使用算法版权获取服务主要包括 2 个步骤：1)鉴权. 判别用户是否具有算法版权. 2)计算. 智能算法调用计算资源来完成服务. 重要的是，鉴权事件是一次性发生或者时域间歇性发生的. 直观来讲，版权提供商验证用户是否具有算法权限的过程是在时域上离散发生的，并不会一直保持鉴权过程. 这意味着可以使用一个版权为多个用户时分交替授予服务权限. 基于此发现，本文提出了互联服务体系的首要设计——版权资源服务化，它将有限版权资源通过时分复用的方式划分为多个时域，独立授权服务单元(下文简称服务单元)，以向更多用户提供无差别甚至更优的算法服务.

但版权资源服务化的实现却面临一个重大挑战：在现有的版权管理体系下，版权资源与用户计算资源一对一高度耦合，无法实现版权在多用户之间的动态流动使用. 为了解决这一难题，我们将多种应用的多个智能算法版权在服务层实现集中式管理，提出智能算法版权管理系统 AIRM. AIRM 系统的约束条件包括算法版权服务响应时间上限 $T_{\max}$、版权数量 N、硬件算力资源 C，在此约束下，服务单元优化目标 S 可定义为：

$$S = \max \sum_{i=1}^{N} f_S(A_i), \forall f_T(S_i) \leqslant T_{\max}, \sum f_C(S) \leqslant C,$$

其中，$f_S(A_i)$表示给定的第 i 个版权 A_i 的可扩增的服务单元数量，$f_T(S_i)$表示某个服务单元的响应时间，$f_C(S)$表示所有服务单元所需的算力资源. 综上所述，本文所提出的 AIRM 系统旨在通过对 N 个智能算法版权的细粒度管理提高有限版权资源的并发服务能力，在 4.2 节深入分析了本文算法用例的服务能力性能增益.

AIRM 还包括版权存储器和服务单元调度器 2 个模块(如图 2 所示)，前者存储了来自多个算法提供商的算法版权，后者对算法版权的细粒度进行调度.

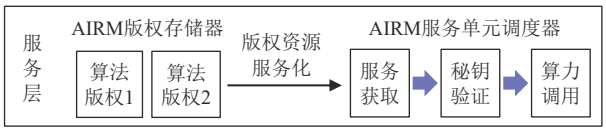

Fig. 2 Design of AIRM module

图 2 AIRM 模块设计

以视频监控中人脸年龄识别应用为例，将多个人脸年龄识别算法版权逻辑定义为版权集合$A=\{A_1, A_2, \cdots, A_n\}$，并存储在版权存储器中. 继而，AIRM通过版权资源服务化技术将A转换为服务单元集合S，存储在服务单元调度器. 不同于传统版权管理系统，AIRM服务单元不会按照时域独享算法版权，而是在定期的鉴权时刻动态请求版权资源来获取算法使用权限. 如图3所示，假如一个人脸年龄算法版权在满足每个用户始终保持鉴权成功的前提下可以通过时域切分为j个时分复用的服务单元，则$S=\{S_{1,1}, S_{2,2}, \cdots, S_{i,j}\}$，其中$A_i$为$S_i=\{S_{i,1}, S_{i,2}, \cdots, S_{i,j}\}$提供鉴权服务，该过程的具体实现方法如算法1所示.

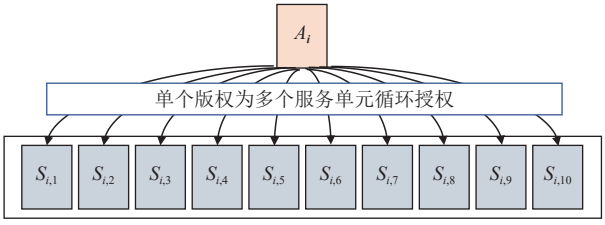

Fig. 3 Cyclic authorization for service units in AIRM

图3 AIRM的服务单元循环授权

算法1. 版权资源按需时域划分方法.

输入：智能算法版权集合$A=\{A_1, A_2, \cdots, A_n\}$，每个算法版权的鉴权间隔集合$T=\{T_{A_1}, T_{A_2}, \cdots, T_{A_n}\}$，每个算法服务用户鉴权时间需求$t_i$，用户服务状态$v_j$；

输出：按需服务单元集合S.

① $S\leftarrow\{\}$，$i\leftarrow1$；

② $j\leftarrow0$，$S_i\leftarrow\{\}$，$size=T_{A_i}/t_i$，$i<n$；

③ while v_j is True and $j<size$

/* v_j is True 代表用户服务处于激活态，未被释放*/

④ S_i. append($S_{i,j}$)；

/*将用户j对应的服务单元添加到S_i集合中*/

⑤ S. append($S_{i,}$)；

/*将S_i服务单元集添加到S中*/

⑥ j++；

⑦ end while

⑧ i++；

⑨ return S.

/*返回A的时域划分结果S*/

具体来说，AIRM不直接让用户通过算法版权获取算法应用能力，而是自定义时分密钥，将每个算法版权拆解成更多的服务单元，设置授权时间片来循环，为更多的用户提供服务. 考虑到不同算法版权的鉴权间隔要求存在明显差异，我们提出了版权资源按需时域的划分方法. 详细地说，依据每个算法具体的鉴权时间间隔T将版权资源按用户所要求的鉴权时间需求t定制化切分为T/t个服务单元. 举例来说，算法版权A_1需要在10 s内为n个用户循环鉴权，若每个用户鉴权时间为1 s，则n=10/1=10；算法版权A_2需要在5 s内为m个用户循环鉴权，若每个用户鉴权时间为0.25 s，则m=5/0.25=20. 因此，在流程上，对于A中的每个智能算法版权A_i，为其建立服务单元映射集S_i，同时获取用户要求的鉴权时间间隔T_{A_i}. 按照按需计算的服务单元数量上限逐渐向每个S_i中增加服务单元，并将其直接分配给用户使用. AIRM也会持续更新S中所有服务单元的服务状态(已完成服务、正在服务、待激活等)，以便用户释放的服务单元被及时分配给新用户使用.

AIRM对版权资源的时域划分是动态调整的. 对于不同的算法版权，其切分的版权资源时隙由其用户端所要求的应用服务授权，间隔进行定制化切分. 在这一机制下，每种算法版权可以为用户提供连续的算法服务，即使出现偶然的版权资源响应超时，AIRM也会及时将用户请求切换至新的服务单元来重新响应. 除此之外，切分好的版权授权时隙会定期通过在用户侧收集到的服务日志进行动态更新，更新周期一般为5~7天，在本文实验中介绍的人脸属性识别算法示例的授权时隙分别为40 ms.

对于每个算法版权，AIRM根据用户鉴权频率确定其时域服务单元划分个数，每个算法版权的授权过程由AIRM平台统一调度. 继而，AIRM为每个服务单元定义二级服务密钥，它用于服务单元所属用户获取算法版权的时分授权，从而获得时域连续的算法服务. 除此之外，由于用户本身网络环境的差异，其对算法应用的调用频率存在差异，所以算法版权的鉴权时间差存在一定浮动. 集中式的版权资源服务化管理方式可以协调多个算法版权在广泛的用户中按需流动，而不是简单地将一个版权时域分发给固定的j个用户. 这也是本文提出的“共享式”AIRM智能版权管理系统与传统“买断式”DRM系统的本质区别. 因此，本文对智能算法资源实现了更细粒度的服务级调度而非传统DRM低效的版权级调度.

随着边缘云计算技术的快速发展，视频数据的云端集中式处理已逐渐成为一种流行的、轻便的解决方案. 一方面，新兴的视频/图像采集器有较强的压缩和抽帧能力，所产生的数据流量较小甚至不足1 Mbps，不会对网络产生较大压力；另一方面，本文所提出的AIRM系统只会在云端处理音视频数据，处

理结果会以 HTTP 消息进行通信，数据量小不会对网络传输带来额外压力.

3.2 流动性算力网络

3.1 节所介绍的基于版权资源服务化的 AIRM 系统在服务层实现了有限数量版权的时域服务化切分，以提高多用户并行服务能力，但是此过程要求更灵活的服务计算资源调度技术. 因此，我们为动态互联服务体系的资源层设计了基于流动性结构的算力网络. 现有的弹性资源调度方案只是单纯的资源层，对物理传输、计算资源的弹性分配缺乏与上层应用版权服务的深度结合. 与其不同的是，AIRM 通过细粒度的物理资源虚拟化技术将云端物理资源细粒度切分并动态映射到版权服务单元以获取算力支持. 特别地，本文的流动性算力网络体现在多个资源层算力单元在多个应用层服务单元之间动态映射，进而提供动态算力支撑，而不是简单的算力单元与服务单元一对一强绑定，这有利于在用户量较小时也可以利用闲置的算力资源，以提高用户服务响应时间. 更近一步，本文所提出的 AIRM 系统还支持跨平台、跨域的流动，不会局限于局域网络与单一平台.

流动性算力网络包含 2 个关键设计：

1）根据上层服务单元需求（服务需求本质上等同于传统管理系统中的用户业务需求），将一定额度的云端算力分配给服务单元，并在完成服务后及时回收；

2）将云端物理资源切分为细粒度的算力单元作为资源调度单位，多个算力单元可以再自由组合和交换，为多个服务单元提供流动性计算能力，实现更好的并行计算效率.

为实现流动性算力网络的构建，需要解决 2 个科学问题：

1）如何划分算力单元，并将物理资源进行逻辑划分成可以通过容器技术实现的问题. 划分后的算力单元代表了系统服务层向资源层的最小计算资源调度粒度. 一般来说，更小的算力单元代表更细的计算资源粒度，造成的计算资源浪费也相对更小. 但是，更小的算力单元也需要更复杂的算力网络结构，会增大算力调度的时间开销.

2）如何实现多个算力单元在多个服务之间的并行流动的问题. 算法服务的完成需要多项物理资源，比如磁盘 I/O、GPU 浮点运算等都具有并行服务能力. 但是，如果算力单元和用户服务一对一耦合，则必然导致对单个算力单元内置资源的串行运行，这将导致较高的资源浪费问题，也会消耗服务提供商更多的运营成本，在实际系统中难以实施.

算法 2. 经验驱动的算力单元划分方法.

输入：算法 A_i 拟被调用的次数 J，历史经验中算法 A_i 第 j 次调用的算力资源 $U_{i,j}$，平台当前待划分的算力总量 G；

输出：拟划分的算力列表 S.

① $i\leftarrow 0$, $j\leftarrow 0$, $U_{i,j}\leftarrow$read Log; /*读取历史记录*/

② while $G\neq 0$; /*循环直至无待划分的算力*/

③ $U_i=\sum U_{i,J}/J$; /*算法 i 的平均算力*/

④ $G=G-(U_i/\sum U)\times G$; /*按比扣除算力*/

⑤ $S_i=(1/J)\times(\sum U_{i,J}/\sum U)$; /*单次算力值*/

⑥ S. append(S_i). repeat(J); /*将算法 i 的拟定的算力单元添加到列表中，加入 J 个*/

⑦ i++;

⑧ end while

⑨ return S.

/*返回算力单元列表 S*/

为了解决问题 1），本文设计了经验驱动的虚拟化算力划分方法. 首先从电信的应用服务平台收集了大量智能算法应用请求的运行轨迹，获取各种算法对物理资源的开销要求. 经过聚类分析后发现：这些算法对物理资源的需求存在明显且非常稀疏的分界线，简单的算法应用对物理资源要求较低，比如视频人脸属性（性别、年龄）的识别；而复杂的算法应用对物理资源的要求较高，比如监控视频中人脸相似性分析. 所以，我们根据算法在历史经验中的算力需求，将物理资源划分为不同粒度的算力单元，使其可以保证同时满足简单算法和复杂算法服务的算力要求. 算力单元的划分粒度根据应用平台用户使用情况周期性动态调整，其具体方法如算法 2 所示. 算力单元的划分方法将根据文中算法 2 输出的算力列表 S 进行，具体地，根据每个算法的历史调用记录及当前的可用算力，对算力列表 S 进行动态更新. 在每次的算力划分中，会将可用待划分算力 G 分配给每一种算法，且每种算法的单次调用算力相同. 系统首先计算历史经验中智能算法 A_i 的平均算力，此后从 G 中扣除智能算法 A_i 被分配到的算力值，该值由记录中 A_i 的算力 U_i 占总记录算力 $\sum U$ 比例而得，即扣除值为 $(U_i/\sum U)\times G$. 此后计算单次算力 S_i 并按调用次数添加到列表 S 中.

为了解决问题 2），本文设计了逻辑资源重组的算力合并方法. 一般来说，多个算力单元是物理非邻接的，无法实现在真实物理层面的资源合并，但是在逻辑上可以合并多个算力单元. 如图 4 所示，将多个

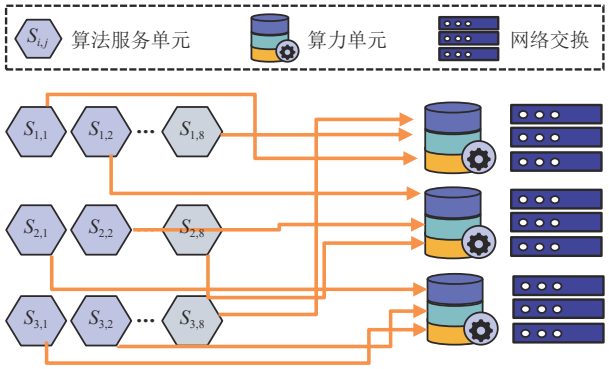

Fig. 4 Resource merging of multiple computation units

图 4 多算力单元的资源合并

算法服务单元的计算任务进行拆解(操作系统级别的并行化计算),并分发至逻辑可合并的多个算力单元以最大化并行处理,从而极大降低服务单元计算任务的处理时间,并将在 4.2 节进行相关实验验证.

当算力资源不足时,服务请求响应时间将增长,若响应时间超过该服务所设阈值,判定为 AIRM 系统处于过载状态. 在此种情况下,一方面,AIRM 会即刻触发服务请求排队机制,此机制采用先进先出(FIFO)策略来响应用户请求. 另一方面,AIRM 会快速扩容资源层算力服务单元,来消除短期的服务请求过载问题,恢复服务的响应时间在正常范围内. 恢复所需的具体时长将受到平台空闲算力资源、网络情况等因素的综合影响,但是尽量保证在毫秒级别达到稳定状态.

4 实 验

本节将介绍视频云网平台中,版权资源服务化与流动性算力网络性能的评估方法及相关实验结果. 将使用并发用户服务量、算法版权响应时间 2 个性能指标,在算力资源变化的条件下评估视频云网平台中不同智能算法的服务性能. 此外,针对不同应用场景中对异构算法服务串行处理的需求,我们测试了组合算法的服务能力与响应时间.

4.1 实验方法

1)实验环境部署. 实验所涉及的版权服务与算力网络环境部署在中电信数智科技视觉智联平台运行,其基于中国电信云计算基础架构,系统服务均部署在云平台虚拟化后的虚拟机或容器中,包括权限认证服务集群、算法存储服务集群以及邦联平台计算资源池 3 个组成部分. 其中,邦联平台资源池提供计算资源以实现诸如 AIRM 管理器、AIRM 容器以及其上层应用的负载. 由虚拟化的节点支撑系统服务的 3 个组成部分,每个虚拟机节点的配置为 Intel Xeon 8 核 CPU、NVIDIA Tesla T4 GPU 与 24 GB RAM 内存,针对计算资源池的场景,本文将这些计算资源划分为 200 个算力单元以组成算力网络并运行多个算法版权实例. 实验过程中,由客户端向平台提交算法服务请求,客户端配置为 Intel Core i5-1035G4 CPU、16 GB RAM 内存. 实验网络环境包括中国电信千兆网与教育网(CERNET)千兆网. 实验过程中算法服务提供 HTTPS 接口,通信报文以 JSON 的格式进行传输,加密方式采用 RSA 签名,以供客户端提交请求和接收算法执行结果.

2)智能视频应用. 本文所设计的版权管理方法可适用于任一类智能算法,例如人脸识别、车辆识别、语音识别等. 本文以人脸识别计算应用为例开展实验. 将视频监控中的人脸图片帧用于人脸识别. 具体包括人脸属性识别与人脸相似度计算 2 个算法,人脸属性识别用于系统并发服务性能的评估,算法的具体模型均由中电信数智科技视觉智联平台提供. 此外,多种算法还会被同时调用以进行算法的组合适配实验.

3)模型输入. 人脸识别应用的输入图像由线上随机抓取或在实验场地拍摄而获得,包括 25 张人脸图像,图片大小为 14~147KB,在实验过程中不断随机抽取图像并提交请求. 由于本文的研究问题在于算力网络与其版权流动、服务能力,故不关注算法推理结果的准确性,无需涵盖标签的数据集.

4)算法调用过程. 用户选择一段视频/图片上传至系统平台,并且选择调用智能算法. 这一应用请求会经由应用层接口下发至服务层,由 AIRM 创建人脸识别服务单元接受用户数据,同时调用资源层算力单元计算识别结果,此结果会被服务层向上传递回应用层展示给用户. 在此过程中,应用层按照 API 规范向其发送请求,包括算法服务码、服务密钥、RSA 签名等. 在平台完成服务计算后,将识别结果回传至用户端,内容包括结果编码、结果描述、服务应答时间戳等.

5)对比方法. 在 DRM 对比实验中,根据算法复杂度为每种算法版权配置不同的算力,实验为每个人脸属性算法版权配置 35 GFLOPS 算力,而为每个人脸相似度算法版权配置 43 GFLOPS 算力,每个算法版权可服务单一容器内的服务请求.

6)性能评估指标. 本文所采用的 4 个主要评估指标为:

①并发服务量. 算法平台在一定算力条件和服务种类的前提下，每秒内可负载的用户最大请求数量由成功返回结果的请求数量除以总耗时计算得出.

②并发服务能力增益. 智能算法版权管理系统并发服务量与传统方案的并发服务量之比.

③响应时间. 算法平台收到了用户请求后，调用算法版权和硬件资源处理该请求的完成时间. 注意，本处强调的响应时间不包括用户请求在网络链路上的传输时间.

④智能算法版权组合适配效率. 在 2 种智能算法组合串行服务场景下，系统并发服务量与增益.

4.2 系统并发服务能力增益

为了评估系统的版权资源服务化能力，本文将量化系统的并发服务能力增益，即比较 AIRM 与传统 DRM 的服务能力差异. 本实验将并发服务量，即最大负载容量作为基本指标用以后续评估. 实验中面临的变量是计算资源，在有限的资源下算法平台可同时处理用户请求的数量有限，此外不同的算法对硬件资源的开销也存在区别. 在本文中，固定实验中调用的算法为人脸属性识别，它是一种常见的算法应用. 用户可以将一张图片发送至算法平台，算法平台将调用版权、执行算法，输出该照片是否存在人脸，以及人脸中的一些必要属性，比如年龄估算、性别分析等. 为了与传统的 DRM 版权管理方案进行对比，在实验中设定为每个算法版权配置 35 GFLOPS 算力(FP16).

图 5 展示了在人脸属性识别应用中，本文所提出的 AIRM 系统与传统 DRM 在不同的算力支撑下并发用户服务量的对比结果. 由于版权“买断式”的缺陷，DRM 的服务量仅能随着算力的扩大而缓慢增长，即每 35 GFLOPS 增加一个服务量. 相较而言，AIRM 得益于版权资源的高效分发，实现了更高量级的并发服务量. 例如，当算力为 350 GFLOPS(FP16)时，DRM 的服务量为 10，AIRM 的并发服务量则可达到 185；

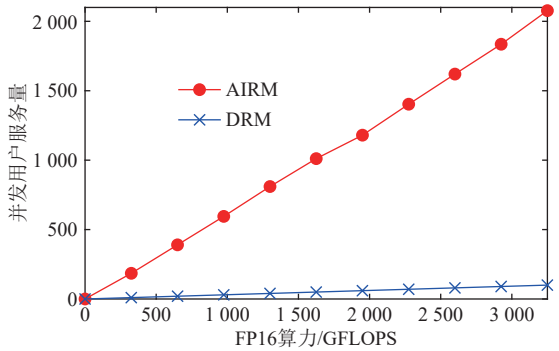

Fig. 5 User service capacity of AIRM and conventional DRM

图 5 AIRM 与传统 DRM 用户服务量

当算力达到 3500 GFLOPS(FP16)时，DRM 的服务量为 100，AIRM 的并发服务量可以扩展到 2077.

本文进一步定义了并发服务能力增益指标用以描述 AIRM 并发能力的提升：平台拥有的版权数量固定，在传统版权管理方法或 AIRM 方法下，分别能够支撑的应用数量之比(1/n, n 为 AIRM 多路应用数量)，即系统的并发服务能力增益，也即 DRM 与 AIRM 的服务量之比. 结果表明，在相同的算力资源下，AIRM 平台的最大用户服务数量是传统 DRM 平台的 19.9 倍(平均值)，并且随着算力资源的逐渐上升，AIRM 的优势更加明显.

4.3 智能算法版权响应时间

为了评估 AIRM 在版权分发服务时延上的优化效果，本文对比了 AIRM 与传统 DRM 的算法版权响应时间. 根据中国电信视频云平台的业务量，在设置不同的版权数量、不同算力的情况下，分别记录每个应用从收到算法请求到生成算法实例的响应时间. 版权与算力的关联与 4.2 节中并发服务能力增益评估时的设定一致，每个算法版权配置 35 GFLOPS 算力(FP16). 本组实验中 DRM 版权管理方案与 AIRM 方案均服务了 20 路用户请求. 将系统处理总时长与算法请求数量进行相除，即为该次实验的请求平均响应时间. 其中系统处理总时长遵循中国电信视频分析平台授权管理模块的处理时长，而不包括实验过程中图像数据上传和下载时间，因其会受到公网网络波动的影响.

图 6 展示了 AIRM 与传统 DRM 在不同的算力资源下用户请求响应时间的对比，DRM 的响应时间均在 0.49 s 左右，而 AIRM 的响应时间随着算力的增长从 0.49 s 逐渐降至 0.40 s，降低算法版权响应时间 18.36%. 该结果表明：随着算力资源的增长，AIRM 平台处理相同数量用户请求的响应时间逐渐降低，而 DRM 平台由于版权服务过程冗余，多个算法版权无法流动，即使算力资源增长也不会对响应时间产生增益. 此外，算力资源对响应时间的增益并不是无限的，随着用户请求被充分分发至多个算法版权，响应时间将逐渐收敛，不再继续下降. 在响应 20 路用户请求，本实验的算力和集群部署条件下，AIRM 的响应时间大致收敛在 0.40 s. 我们猜测，该收敛时间会随着服务数量、算力部署情况的变化而产生不同结果. 此外，根据算力与 AIRM 响应时间的关系，可依据响应时间来评判系统是否处于过载状态，评判阈值一般被设定为传统 DRM 系统理论响应时间(本实验即为图 6 中的时间上限 0.49 s)，当 AIRM 响应时间超过

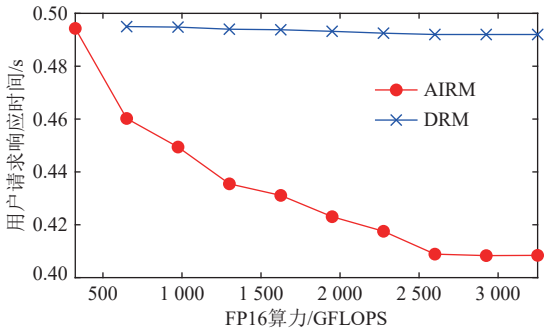

Fig. 6 User request response time of AIRM and conventional DRM

图 6 AIRM 与传统 DRM 下的用户请求响应时间

这一阈值时，判定系统处于过载状态，将进行算力单元扩容，此过程会在毫秒级别内处理完成，尽早恢复到稳定状态.

4.4 智能算法版权组合适配效率对比

在视频云的相关应用中，广泛存在着算法组合的需求. 算法版权的组合适配需要版权资源服务化与算力网络的共同支持，从而实现不同的算法版权与算力资源在系统内的分发流动. 典型的案例是在视频监控人脸识别的需求中，需要同时利用到人脸属性识别(简称“算法-1”)与人脸相似度计算(简称“算法-2”)2 种智能算法. 大多数场景中，“算法-1”对计算资源需求小，版权价格便宜，适合长时间的待机监控检测；而“算法-2”版权价格昂贵且计算资源消耗高. 为了实现减小开销的目的，AIRM 仅在“算法-1”检测到必要目标时调用“算法-2”，以此降低版权与算力负载. 通过 AIRM 实现“算法-1”与“算法-2”的组合适配，可以降低应用成本，在有限的版权资源数量下实现更大的应用容量.

本实验将评估 AIRM 在算法版权组合适配应用场景中的效率. 在本节实验中为了使算力与版权的对应更符合应用场景设定，在 DRM 对比实验中设定为每个人脸属性算法版权配置 35 GFLOPS 算力，而为每个人脸相似度算法版权配置 43 GFLOPS 算力. 实验将分多组不同的算力进行，实验过程中由客户端向 AIRM 连续发出请求. 客户端首先批量进行人脸属性算法请求，持续 0.5 s；再批量请求人脸相似度计算，持续 0.5 s，以此循环直至达到预置的请求数量.

在算力为 350~3500 GFLOPS 变化时，AIRM 的组合服务能力与 DRM 服务能力的服务量对比如表 1 所示. 从实验结果中可以观察到 AIRM 相比于 DRM 仍然具有显著的服务能力提升效果，例如在算力为 350 GFLOPS 时，DRM 可以完成 4 组算法服务，而

Table 1 AIRM Algorithm Combination Service Capability Gain

表 1 AIRM 算法组合服务能力增益

GPU 算力/GFLOPS	DRM 组合服务量	AIRM 组合服务量	组合并发服务能力增益	平均服务能力增益
350	4	30	7.50x	8.07x
700	8	62	7.75x	
1 050	12	95	7.92x	
1 400	16	128	8.00x	
1 750	20	161	8.05x	
2 100	24	195	8.13x	
2 450	28	230	8.21x	
2 800	32	266	8.31x	
3 150	36	301	8.36x	
3 500	40	337	8.43x	

AIRM 可以完成 30 组算法服务，即实现了 7.5 倍的提升. 在不同算力条件下，AIRM 组合服务量平均有 8 倍的提升，表明了智能版权管理系统在算法组合适配条件下的优越性能.

对比 4.2 节中仅进行人脸属性识别时的服务能力，在算法组合时相同算力的服务能力会有所下降，例如表 1 中算力为 350 GFLOPS 时，AIRM 组合服务量为 30，而图 5 中该算力下 AIRM 服务量为 185. 这是由算法组合的资源负载提升导致，在相同的时间内处理 2 种算法的请求，会导致并发服务量下降，尤其是人脸相似度计算(“算法-2”)对算力资源的要求更高.

此外，对比表 1 与 4.2 节中的服务能力增益，算法组合的 AIRM 并发服务能力增益为 8 倍，而单一人脸属性识别的服务能力增益约为 20 倍. 这种增益能力的变化是由于 AIRM 对不同计算复杂度的算法并行能力不同，计算复杂度低的算法会具备更强的版权并行分发能力；而组合算法提高了计算复杂度，致使 AIRM 的并发增益能力衰减. 此外，目前的实验设定中 AIRM 将 2 种算法分 2 批次处理，2 个算法提交的请求数量相同，这是在以最大的业务容量下做出的假设，实际情况中“算法-2”的请求量将小于“算法-1”的请求量. 若在实验过程中进一步优化 2 个批次中“算法-1”与“算法-2”的提交请求数量比例，AIRM 的算法组合服务能力还将进一步得到提升.

5 总结与展望

智能视频云技术正在加速发展，视频云服务对

云网平台中的算力资源提出了非常高的要求，针对大量的智能算法管理问题，本文提出了“算法即服务”理念，建立了新型的“服务—算法—资源”动态互联服务体系，有效解决算法快速迭代、应用需求时变与智能算法版权固化管理的矛盾. 本文还提出智能算法版权管理系统，设计了细粒度版权资源服务化与流动性算法网络化方法，实现智能算法版权管理相关功能模块. 特别地，在中国电信视频分析平台授权管理模块的实际部署与规模化分析验证了本文所设计方法的先进性.

针对本文所提出的智能算法版权管理系统，未来可以在2个方面继续开展研究：1)面向大规模商用系统的智能版权低时延调度方法. 在实际系统中，不同的应用平台对算法请求具有不同的响应时间要求，如何设计具有实时可扩展能力的智能算法版权管理方法，为未来的研究工作之一. 2)智能算法版权管理系统中的安全问题. 在系统实际使用过程中，如何实现用户终端与系统的安全接入、安全连接与可靠性恢复，将在未来工作中展开深入研究.

作者贡献声明：张欢欢提出了算法主要思想；安聪凯设计了算法实现方法；赵朗程完成了实验方案；周安福与马华东负责整个项目研究思路，提出指导性意见并修改论文；袁艺和曹宁提出了在实际系统中的应用思路及建设方案，并提供实验平台支持应用验证.

参 考 文 献

[1] Zhang Qingyang, Sun Hui, Wu Xiaopei, et al. Edge video analytics for public safety: A review[J]. Proceedings of the IEEE, 2019, 107(8): 1675−1696

[2] Hussain T, Muhammad K, Ullah A, et al. Cloud-assisted multiview video summarization using CNN and bidirectional LSTM[J]. IEEE Transactions on Industrial Informatics, 2019, 16(1): 77−86

[3] Liu Qiong, Safavi-Naini R, Sheppard N P. Digital rights management for content distribution[C] //Proc of the Conf in Research and Practice in Information Technology Series. New York: ACM, 2003, 34: 49−58

[4] Petrlic R, Sorge C. Privacy-preserving DRM for cloud computing[C] //Proc of the 26th Int Conf on Advanced Information Networking and Applications Workshops. Piscataway, NJ: IEEE, 2012: 1286−1291

[5] Lee H, Park S, Seo C, et al. DRM cloud framework to support heterogeneous digital rights management systems[J]. Multimedia Tools and Applications, 2016, 75: 14089−14109

[6] Fan Xuefeng, Zhou Xiaoyi, Zhu Bingbing, et al. Survey of copyright protection schemes based on DNN model[J]. Journal of Computer Research and Development, 2022, 59(05): 953−977 (in Chinese)
(樊雪峰, 周晓谊, 朱冰冰, 等. 深度神经网络模型版权保护方案综述[J]. 计算机研究与发展, 2022, 59(05): 953−977)

[7] Fan Lixin, Ng K W, Chan C S. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks[J]. Advances in Neural Information Processing Systems, 2019: 4716−4725

[8] Yang Peng, Lao Yingjie, Li Ping. Robust watermarking for deep neural networks via bi-level optimization[C] //Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 14841−14850

[9] Zhang Jie, Chen Dongdong, Liao Jing, et al. Passport-aware normalization for deep model protection[J]. Advances in Neural Information Processing Systems, 2020, 33: 22619−22628

[10] Chen Huili, Rouhani B D, Fu Cheng, et al. Deepmarks: A secure fingerprinting framework for digital rights management of deep learning models[C] //Proc of the 2019 Int Conf on Multimedia Retrieval. New York: ACM, 2019: 105−113

[11] Ribeiro M, Grolinger K, Capretz M A M. Mlaas: Machine learning as a service[C] //Proc of the 2015 IEEE 14th Int Conf on Machine Learning and Applications (ICMLA). Piscataway, NJ: IEEE, 2015: 896−902

[12] Sapio A, Canini M, Ho C Y, et al. Scaling distributed machine learning with In-Network aggregation[C] //Proc of the 18th USENIX Symp on Networked Systems Design and Implementation (NSDI 21). Berkeley, CA: USENIX Association, 2021: 785−808

[13] Weng Qizhen, Xiao Wencong, Yu Yinghao, et al. MLaaS in the Wild: Workload analysis and scheduling in large-scale heterogeneous GPU clusters[C] //Proc of the 19th USENIX Symp on Networked Systems Design and Implementation (NSDI 22). Berkeley, CA: USENIX Association, 2022: 945−960

[14] Xiao Wencong, Ren Shiru, Li Yong, et al. AntMan: Dynamic scaling on GPU clusters for deep learning[C] //Proc of the 14th USENIX Symp on Operating Systems Design and Implementation (OSDI 20). Berkeley, CA: USENIX Association, 2020: 533−548

[15] Zhang Chengliang, Yu Minchen, Wang Wei, et al. MArk: Exploiting cloud services for cost-effective, SLO-aware machine learning inference serving[C] //Proc of the 2019 USENIX Annual Technical Conf (USENIX ATC 19). Berkeley, CA: USENIX Association, 2019: 1049−1062

[16] Zhao Hanyu, Han Zhenhua, Yang Zhi, et al. HiveD: Sharing a GPU cluster for deep learning with guarantees[C] //Proc of the14th USENIX Symp on Operating Systems Design and Implementation (OSDI 20). Berkeley, CA: USENIX Association, 2020: 515−532

[17] Han Ying. Network copyright protection in cloud computing environment and countermeasures[J]. China Publishing Journal, 2012(10): 54−56 (in Chinese)
(韩缨. 云计算环境下网络版权保护问题和应对策略[J]. 中国出版, 2012(10): 54−56)

[18] Wang Jing, Huang Chuanhe, Wang Jinhai. An access control mechanism with dynamic privilege for cloud stroage[J]. Journal of Computer Research and Development, 2016, 53(04): 904−920 (in

Chinese)
(王晶, 黄传河, 王金海. 一种面向云存储的动态授权访问控制机制[J]. 计算机研究与发展, 2016, 53(04): 904−920)
[19] Huang Qinlong, Ma Zhaofeng, Fu Jingyi, et al. Privacy-preserving digital rights management scheme in cloud computing[J]. Journal on Communications, 2014, 35(02): 95−103 (in Chinese)
(黄勤龙, 马兆丰, 傅镜艺, 等. 云计算环境中支持隐私保护的数字版权保护方案[J]. 通信学报, 2014, 35(02): 95−103)
[20] Poddar R, Ananthanarayanan G, Setty S, et al. Visor: Privacy-preserving video analytics as a cloud service[C] //Proc of the 29th USENIX Security Symp (USENIX Security 20). Berkeley, CA: USENIX Association, 2020: 1039−1056
[21] Petrlic R. Proxy re-encryption in a privacy-preserving cloud computing DRM scheme[C] //Proc of the Int Symp on Cyberspace Safety and Security. Berlin: Springer, 2012: 194−211
[22] Liu Shaobin, Ma Jianfeng, Feng Xiaoqing. Transparent access and integration of heterogeneous encrypted database in hybrid cloud environment[C] //Proc of the 2019 IEEE Int Conf on Communications (ICC). Piscataway, NJ: IEEE, 2019: 1−6

Zhang Huanhuan, born in 1994.Postdoctoral researcher. Her main research interests include IoT sensing, mobile computing, and low-latency video streaming optimization.
张欢欢, 1994 年生. 博士后. 主要研究方向为物联网感知、移动计算、低时延视频流优化.

An Congkai, born in 1994. PhD candidate. His main research interests include network congestion control and video streaming transport optimization.
安聪凯, 1994 年生. 博士研究生. 主要研究方向包括网络拥塞控制、视频流传输优化.

Zhao Langcheng, born in 1997. PhD candidate. His main research interests include IoT sensing and mobile computing.
赵朗程, 1997 年生. 博士研究生. 主要研究方向为物联网感知、移动计算.

Zhou Anfu, born in 1981. Professor. His main research interests include IoT sensing and mobile computing.
周安福, 1981 年生. 教授. 主要研究方向为物联网感知、移动计算.

Ma Huadong, born in 1964. Professor. His main research interests include IoT, sensor network and multimedia computing.
马华东, 1964 年生. 教授. 主要研究方向为物联网、传感网、多媒体计算.

Yuan Yi, born in 1994. Master. Her main research interests include intelligent video cloud and video network.
袁　艺, 1994 年生. 硕士. 主要研究方向为智能视频云、视频网络.

Cao Ning, born in 1976. Master Candidate. His main research interests include intelligent video cloud and video IoT.
曹　宁, 1976 年生. 硕士研究生. 主要研究方向包括智能视频云、视联网.