



虚拟实时服务:构架、算法及演进路径

陈 巍

(清华大学电子工程系 北京 100084)

摘 要:提出了一种基于虚拟实时服务的协议构架,可有效发掘用户提出通信请求之前的“负时间轴”上的空闲时频资源或可再生能源,将用户可能感兴趣的内容以多播的方式预先推送到终端的存储器中,从而以非实时的通信提供虚拟实时服务,在确保服务质量的前提下大幅度提升频谱和能量效率。对终端缓存内容的选择算法进行了标准化,从而允许基站在本地预测用户已缓存的内容,并判断当前是否应发起推送和推送哪条内容;提出了虚拟实时服务涉及的若干关键算法,也给出了虚拟实时服务的演进途径及主要开放问题,旨在为相关新兴产业的技术发展提供参考。

关键词:虚拟实时服务;多元协同通信;透明多播推送;认知频谱共享;能量收割;频谱效率;能量效率

doi: 10.11959/j.issn.1000-0801.2015129

Virtual Realtime Services: Infrastructure, Algorithms and Evolution Path

Chen Wei

(Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: A virtual real-time service infrastructure for future wireless access, which is capable of efficiently using the idle timeslots or renewable energy distributed in the “negative time-axis” before users’ transmission requests was proposed. By using these idle timeslots, the information content in which users might be interested can be pushed to the end devices’ buffers in a multicasting manner before a user requests the content. The end devices’ algorithm which determines which content should be buffered was standardized. Hence, the base-station may locally predict what have been buffered in the users’ devices, and determine whether it should push and which content to transmit. Meanwhile, a number of key algorithms for virtual real-time service were proposed and analyzed, an evolution path toward virtual real-time service along with its major problems was also presented. Chen’s program was proposed to guide the development of the virtual real-time service related emerging industry and technologies.

Key words: virtual real-time service, multi-unit collaborative communication, transparent multicast push, cognitive spectrum reuse, energy harvesting, spectral efficiency, energy efficiency

收稿日期:2015-05-06;修回日期:2015-05-15

基金项目:“密集立体覆盖移动通信的基础理论与方法”国家重点基础研究发展计划“973”计划青年科学家专题基金资助项目(No.2013CB336600),国家自然科学基金优秀青年科学基金资助项目(No.61322111)

Foundation Items: National Basic Research Program of China (973 Program) (No. 2013CB336600), National Natural Science Foundation of China, Excellent Young Investigator Award (No. 61322111)

论文引用格式:陈巍. 虚拟实时服务:构架、算法及演进路径. 电信科学, 2015129

Chen W. Virtual realtime services: infrastructure, algorithms and evolution path. Telecommunications Science, 2015129



1 引言

不断提升频谱效率和能量效率,是无线网络演化进程中人们矢志不渝的追求^[1-3]。特别是,近年来随着移动互联网的普及,各类新兴的移动互联服务如雨后春笋般涌现,导致指数增长的无线业务量和逐渐分配殆尽的无线频谱资源之间的矛盾日益凸显。与此同时,随着调制阶数的提升、基站密度的增加,无线网络也逐渐成为高能耗企业的信息服务基础设施,社会各界对其节能化、绿色化的呼声日益高涨。

在这一应用背景下,研究人员从如下3个层面着手解决无线网络演进需求与频谱、能量效率限制之间的矛盾。第一,研发先进的物理层传输技术,将已有频谱和常规能源的利用效率不断提升到其理论性能极限,代表性技术包括OFDM、LDPC(low density parity check code,低密度奇偶校验码)以及近年来兴起的物理层多播等技术^[4-6];第二,拓展无线资源及其空间复用维度,寻求新的用于无线通信的频谱资源、复用维度,以缓解常用频谱资源和复用维度的效率压力,代表性技术包括毫米波通信、无线光通信、MIMO、干扰校准、极化复用和立体复用等技术^[7-15];第三,在现有的固定频率分配、常规能源支持之外,发掘空闲的用于无线传输的频谱资源、可再生能源,代表性技术为认知无线电技术^[16]、能量收割通信技术^[17]。

本文着重探讨了如何综合发挥认知无线电、能量收割和物理层多播技术的联合增益,以提升能量和频谱效率。回顾已有的研究工作,上述3种技术虽然有一定的潜在谱效和能效增益,但是在实际应用中的核心困难是——难以

对实时业务提供服务质量保障。具体来说:首先,认知无线电技术利用主系统(授权系统)的空闲时隙进行通信,因此当主系统持续繁忙时,次级用户系统(非授权系统)中等待传输的业务就会有较大的排队时延^[18];其次,能量收割系统利用非平稳到达的可再生能源进行通信,一方面业务排队等待能量的到达会产生额外的时延,另一方面当业务需求量较少时,会导致累积的收割能量超过电池容量而被浪费掉^[19,20];最后,在物理层多播系统中,不同用户对相同内容的索取时刻不同,若要获得多播增益就必须对齐它们的传输时刻,同样会造成部分用户的等待时延^[21]。究其原因,本质上都是具有较高频谱和能量效率的时机与业务需求的时刻不匹配,从而对实时性要求较高的业务来说,无法有效地保障其服务质量,换言之,此类业务无法有效地利用认知无线电、能量收割和物理层多播技术来提升频谱和能量效率。

为了克服上述局限性,在之前研究工作的基础上,提出了一种基于虚拟实时服务的新型服务体制^[22],通过将推送机制和认知无线电、能量收割和物理层多播等技术相结合,能够有效缓地缓解上述业务需求时刻和(频谱、能量)资源供给时刻不匹配的问题。其核心思想是,通过提前预测用户的需求,在用户发起业务请求之前的负时间轴上发起业务推送,这样不仅可以充分利用分布在负时间轴上的大量空闲时隙(包括系统内的空闲时隙和系统外的空闲时隙),提升频谱效率,还有足够的时间可以通过合理时延、对齐不同用户的推送时刻,来获取物理层多播的能效和谱效增益,同时,对于具有能量收割功能辅助的通信系统,能够在更大的时间尺度上收集和利用可再生能源^[23],从而避

首席科学家

陈巍

男,博士,清华大学电子工程系教授、博士生导师(工学、工程)、副系主任、教学委员会主任、清华大学“电子与通信工程”学位评定分委员会副主席,中国通信学会青年工作委员会副主任委员,中国电子学会教育工作委员会青年工作组副主任委员,北京市通信学会青年工作委员会委员,IEEE Transactions on Education、IEEE Wireless Communication Letters、China Communications 等国际期刊编委,“全国五一劳动奖章”、“北京市优秀青年人才”、“北京市优秀教师”荣誉称号获得者,国家重点基础研究发展计划(“973”计划)首席科学家,国家自然科学基金优秀青年科学基金、第十四届霍英东高校青年教师奖一等奖、第十七届茅以升北京青年科技奖获得者,IEEE 高级会员,国家中青年科技创新领军人才计划、教育部新世纪优秀人才支持计划、北京市科技新星计划和清华大学“221”人才计划入选者。主要研究方向为无线通信、信息论、移动互联网。

免了因电池容量不足而导致的能量浪费。上述频谱和能量效率的增益都是在无需付出业务时延的条件下获得的,因此,预先推送可以用非实时的传输,使用户获得虚拟实时的服务体验。

相对于传统通信体制中,由用户请求发起的“拉式通信”,虚拟实时服务将其模式变为“推式通信”,即以空闲频谱资源、收割到的能量以及多播机会来驱动发起推送行为。虚拟实时服务需要计算、存储、通信等各单元之间的多元协同,以消耗存储空间为代价,换取能量和频谱效率的提升。为了适应虚拟实时服务的量化建模,提出了一种新的信息度量单元——内容。回顾不同层次的信息传输系统,它们都有建模时常用的基本单元,在传输过程中不分割讨论,如点到点传输的基本传输单元 bit,通信网的基本传输单元分组或者帧。为讨论虚拟实时服务,定义了一种基本单元——内容。内容可以是一篇文章、一段视频等,其信息量尺度高于 bit 和分组,但是从服务的角度来说,对用户具有一定的不可分割性,必须作为整体推送完成。目前,微信、新闻阅读器应用中的一篇文章、一段视频等就可以近似看作一条内容。

由于虚拟实时服务的推送不是由用户需求发起的,因此内容的传送完成时刻和用户需要此内容的时刻之间会有时间差,在这段时间中内容需要存放在用户缓存中,消耗有限的存储资源;更甚者,若推送的内容根本不是用户所需要的,会导致存储资源的浪费,对提升频谱和能量效率也无贡献^[24,25]。因此,推送并非越早越好、越多越好,而是要准确预测用户的需求,并在合适的时机发起,既不消耗过多存储资源,又使得推送的内容尽量为用户所用。因此,

如何设计实用的虚拟实时服务协议与算法,一方面有效地利用用户终端有限的存储空间,另一方面最大化频谱效率和能量效率的增益,是本文需要解决的问题。

针对上述问题,提出了一种基于虚拟实时服务的实用协议构架,可以“半透明”地使用用户存储终端中的剩余空间,以最小的存储代价获取频谱和能量效率增益。为此,首先提出了一种量化描述用户对内容的需求行为的模型;在此基础上,提出了一种计算当前时刻用户对某个内容的需求度的方法,该用户通过对不同内容的需求度进行比较,决定此刻在有限的缓存区中应保存的内容。基于用户端的缓存区内容保留选择算法,进一步提出了基站端通过预测和反馈,了解用户缓存区占用情况的方法,并在此基础上判断当前需求度最高的内容,决定这一内容是否值得推送。由半定量分析可以看出,在虚拟实时服务中,存在能量效率和频谱效率的基本折中关系,同时,在线的需求学习和预测算法设计,对系统的性能也有重要的影响。

2 虚拟实时服务的基本假设和协议构架

本节首先给出了设计虚拟实时服务协议的基本假设,在此基础上对收发端协议的整体构架进行了介绍。

2.1 假设一——可预测准则

在虚拟实时服务中,系统必须可以预测用户的需求,从而能够在合适的时机推送合适的内容。尽管已有的研究工作对内容被需求的概率已经有刻画,但是本文关注的是用户对某一内容提出需求的时延特性,而不仅仅是是否需求的二元特性。这是因为:即使是被用户需要的内容,如果过早推送而在用户的缓存区内停留太长时间,同样也会造

项目简介

本项目针对高层建筑室内高速、可靠、实时无线覆盖的需求,针对当前移动互联网业务激增和无线通信频谱资源稀缺之间的内在矛盾,探索了通过多域协同、跨层设计、智能认知,大幅度提升无线网络的公平性、实时性和业务容量的方法。在基础理论研究层面,突破了国际上发展 5G 后高层建筑室内移动互联网的若干难题,提出了面向高层建筑内多层频谱复用的理论框架和性能测度,阐明了整体性能极限及逼近方法;在应用基础研究层面,力争占据在 5G 后多层建筑室内移动互联网的技术发展源头和战略制高点,提出、设计和优化了可有效提升层间频谱复用异构无线网络服务质量的多种网络协议与通信算法。

根据信息论基本原理,频谱和能量资源是决定无线通信网络容量的根本因素,而频谱还能换取能效的大幅度提升。在信息论提出之后的 60 多年中,国际上无线通信的研究始终围绕如下两条思路展开:提升已有无线资源的利用率;发掘全新的无线资源或维度。

近年来的工作热点,都可以归入上述两条思路。围绕第一条思路所开展的具体工作包括高阶调制、编码调制(TCM)、逼近 Shannon(香农)界的信道编码(LDPC、Turbo)等传统工作。从 20 世纪 90 年代末期开始通信领域又着眼于第二条思路。



成存储资源的严重浪费。

基于上述考虑,提出一种基于概率论的描述方法。内容以某种概率分布随机产生并到达基站端,用非负整数序号 i 标识一条内容,定义 T_{ik} 为用户 k 对内容 i 提出使用需求的时延,即:内容 i 产生后过了 T_{ik} 时间,被用户 k 要求使用(阅读、观看等), T_{ik} 为非负随机变量,其概率累积函数记为 $F_{ik}(x)=Pr\{T_{ik}\leq x\}$ 。这一分布对于基站和用户双方是已知的。任意一条内容 i ,都有一个简短的标签,使得用户可以大致了解其特征,根据标签决定是否、何时请求该内容,从而影响用户对该内容的需求模式。目前,各新闻或视频网站、微信等均具有类似功能,同时因标签仅有数 bit、类似于超链接,因此其传输、存储的代价均可以忽略。通过学习标签的内容、关键词,系统可以对内容 i 进行归类,并根据对此类内容之前统计的需求时延的经验分布,得到本内容的需求时延的近似分布 $F_{ik}(x)$ 。对此,机器学习、人工智能和社交网络等领域已经有一定的研究基础,有助于更好地进行分类和估计。对用户群体可以进一步归类,例如某类用户具有相同的 $F_{ik}(x)$,则统计学习的收敛速度更快,准确性更高,调度算法更简化。还需强调的是,内容是具有生命周期的,当超过这一周期后即被删除,此时其序号 i 可以被释放出来重复使用,类似于 ARQ (automatic repeat request, 自动重复请求协议) 中帧序号的标记机制。

2.2 假设二——有效性准则

频谱效率和能量效率是通信理论中的两个重要指标,其计算离不开对“有效传输的数据量”的定义。在经典通信理论中,有效传输的数据量即正确接收的数据量,因为此时传输是用户请求发起的,其正确接收的数据对用户必然

是有用的。而在推送模式下,由于用户需求的不确定性,终端正确接收的数据未必被用户使用。因此,需重新定义“有效传输的数据量”,即用户提出传输请求后,能够在其存储单元中找到的数据。基于这一定义,可以进一步刻画虚拟实时服务为频谱效率和能量效率带来的增益。由于虚拟实时服务使用的是空闲频谱,因此只要用户使用了存储单元中的信息内容,就意味着节省了对频谱的使用,所以把频谱效率的增益定义为:单位时间、单位带宽上,用户从终端存储单元读取的数据量总和。需要注意的是,推送是以多播的形式进行的,因此在无线环境中增益巨大。类似地,把能量效率定义为:使用单位能量进行多播推送,能够让用户从终端存储单元中读取的数据量总和。能量效率同样也随着多播组终端数量的提升而提升。当空闲间隙较多时,增大推送频次始终能提升频谱效率(因为空闲频谱是“免费”的),但可能降低能量效率,这种折中关系在后面还会讨论。此外,若采用能量收割技术允许基站使用可再生能源,则这部分能源不计入消耗的能量,即能效计算中的分母。

2.3 假设三——透明化准则

希望用户几乎感觉不到此类服务是“虚拟实时”的,或者说,虚拟实时服务几乎不对通信系统、存储系统带来任何易于察觉的影响,故用户无法区分所接收的信息内容来自于即时的按需传输,还是预先的推送。为了满足这一准则,在协议设计中需遵循如下两点:对内容请求“即时响应后清除”;对终端存储器“透明使用”。

“即时响应后清除”的目的,是为了避免改变用户对网络的使用和对内容的索取行为模式,从而造成系统内部的反馈不稳定。具体地说,当用户提出内容传输请求,而其存

项目简介

为了提升点到点通信的容量,利用多天线资源实现空间维度复用的 MIMO 技术从 1998 年开始成为无线通信的研究热点。最近几年,这种基于多天线思想,发掘空间资源的工作又被进一步推广到中继技术、干扰校准技术以及轨道角动量调制等,持续被学术界关注。然而,信息论和电磁场的交叉研究表明,即使空间资源得以充分开发,无线网络容量从本质上还受制于带宽,无法扩展。因此,国际学术界又开始从频率维度发掘新的频谱资源,典型的技术包括高频段、太赫兹和无线光通信等。此外,近年来通信学界还与物理学进行深入交叉,从电磁场的其他自由度(如极化方向等)需求维度拓展,目前已有工作发表于《Nature》。

本项目中拟开展的研究工作源于第二条思路,希望系统地探索目前国际学术界尚未发现的无线资源维度。在空间维度,利用“平面之外的第 3 个维度”实现空间频谱复用,提出创新的网络构架——“密集立体覆盖”;在时间维度,利用“负时间轴”上的空闲频谱资源,提出创新的服务机制——“虚拟实时服务”。

储单元中没有该内容时,基站应通过实时服务链路向该用户传输请求的内容,这传统网络中基站对用户需求的响应模式完全一致,因此,用户几乎无法分辨所获得的内容是来自于预先推送,还是来自于即时的按需传输,从而感受到虚拟实时服务的透明性。

从之前定义的能量和频谱效率增益的角度来看,希望尽可能减少即时的按需传输,因为这一传输模式既难以利用空闲的时频资源和可再生能源,又没有多播增益,因此频谱和能量效率较低。同时,也不允许用户提出请求后等待内容的多播推送,因为这样不仅会破坏服务质量,也会导致协议设计中的正反馈。

在虚拟实时服务中,推送到终端的内容需暂时缓存在其存储单元内。在“透明使用”模式下,推送内容只能占用空闲的存储空间,一旦空闲存储空间缩小或不能满足推送内容的存储需求,存储单元就必须删除部分已推送的内容。在这一模式下,终端用户感觉不到推送对设备存储空间带来的任何影响,因此是“透明”的。此时,对空闲存储空间的使用类似于认知无线电中次级用户对空闲频谱的使用,不会产生任何额外代价。当然,可以专门划分一部分空间用于存储推送的内容,对于超出存储需求的部分则采用上述透明使用模式,该模式称为半透明存储模式。类似地,“半透明”概念也适用于信道的使用,可以专门划分一部分时频资源,用来支持虚拟实时服务的推送,对于超出这部分时频资源的传输需求,则以次级用户的模式发掘空闲时频资源进行复用,该模式也可称为半透明传输模式。

2.4 虚拟实时服务的协议构架

基于上述3条假设,接下来对虚拟实时服务的协议构

架进行介绍,如图1所示。这一构架最大的特色是基站在本地,利用终端的缓存区内容保留选择算法,结合终端的内容请求信息,预测最能产生推送效用,即在统计意义上用户需求度最大的内容,并决定当前时刻是否发起推送。这一点借鉴了增量调制、差分编码等信源编码的设计思想。

在虚拟实时服务中,终端由存储单元和用户端缓存器的内容选择单元组成,如图1中虚线右边所示。在终端存储单元中,可供虚拟实施服务使用的空间由两部分组成:一部分是确定的划分给此类服务使用的专用空间;另一部分是存储器中未被其他业务占用的空闲空间,其可用存储容量可用随机过程来描述,且此类随机过程多数具有Markov性。两类存储空间总容量决定了缓存内容容量的上限,当存储容量不足时,就需要终端设计一种算法,选择保留哪些内容,丢弃哪些内容,此算法称为用户端的缓存器内容选择算法。该算法根据用户 k 对内容 i 的需求时延分布 $F_{ik}(x)$ 来预测该用户对不同内容的需求度,从而决定不同内容需被保留的优先级。此外,由于 $F_{ik}(x)$ 常常需要通过学习来获得,因此,有时还需辅助一个 $F_{ik}(x)$ 的学习算法模块,为缓存器内容选择算法提供决策参考。最后,假设用户不会重复阅读、观看或使用同一内容,那么,当用户提出传输请求后,若在存储单元内找到了所需的内容,则该内容随即被存储单元删除。

虚拟实时服务的基站构架如图1中虚线左边所示。基站端包含基站的存储单元、推送决策模块和以用户端内容选择算法为核心的本地预测模块。这一构架的设计思想受到了增量调制等信源编码的启发,将接收端的算法放到发送端进行本地预测,结合用户发来的内容请求,基站可以

项目简介

本项目围绕的核心内容是高层建筑室内可重配置异构云无线接入网络。针对两个突破层次(即任务层次,有效支持超高用户密度超感官需求的数据接入需求;理论层次,高层建筑室内立体接入网络的容量极限及逼近方法),提出3个创新方法:可有效适应于环境的立体频谱复用架构;用户自主部署网络的认知协同受控自治;计算嵌入的分布式与精准无线资源管理。

由于室内容量需求巨大,预计2020年业务需求还将提升千倍,而室内业务量的增长超过90%,使得这一工作意义重大。在本项目中聚焦于3个“一”:一个场景,典型高层建筑的室内高速无线接入;一个对象,适应建筑环境的可重配置异构云无线网络;一个问题,高层建筑室内通信的认知协同机理及性能极限。

紧扣适宜高层建筑室内覆盖的无线网络的3个关键特征开展如下研究工作:大规模立体分布式天;可重配置软件定义网络;基于认知的分布式协同。

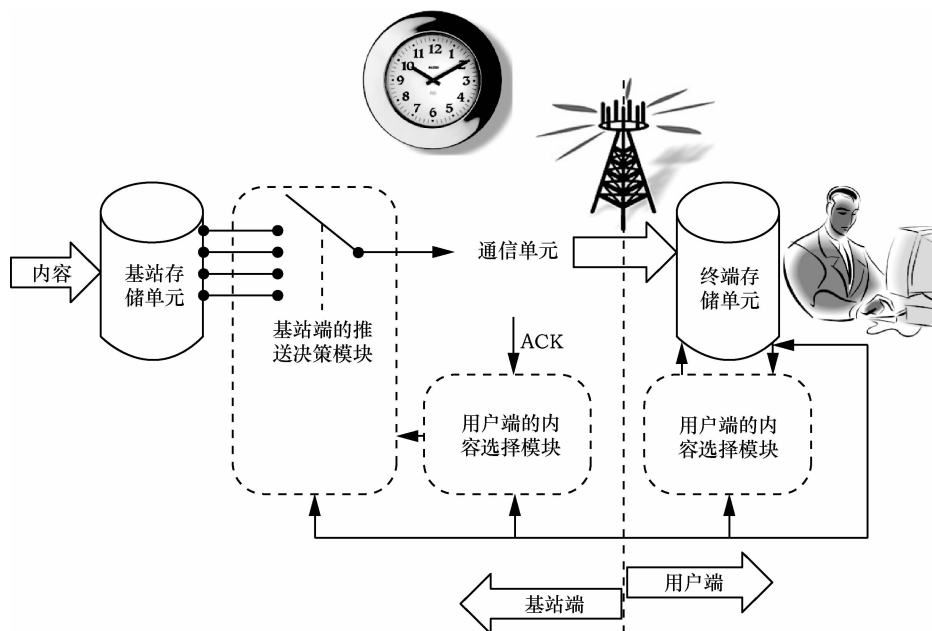


图1 虚拟实时服务的总体协议构架

知道目前在终端的缓存区内有哪些内容,剩余空间还有多大。基于这一本地预测结果,推送决策模块可扣除用户缓存区已有的内容,分析预测每个用户对每个内容的需求度,同时综合考虑用户的剩余空间情况,决定是否发起推送,推送哪条内容。推送内容缓存区负责缓存到达基站的内容,并在其生命周期结束后将其删除,类似于信源编码。建议应对用户端的缓存内容选择算法元建立行业标准,而开放推送内容选择模块的算法设计,留给不同系统自行开发。

最后,介绍一下基站和终端之间的信令交互过程,下行链路主要以推送业务为主,上行链路主要传递用户终端业务请求的信令,无论用户请求的业务是否能够在

本地缓存中找到,都必须向基站发送该请求,以便基站能够更新本地预测的缓存区情况,并统计用户的需求时延概率分布。

从以上论述还可以看出,在虚拟实时服务中,若不划定固定的通信时隙和存储空间,则对用户来说,整个实时服务的虚拟化过程是完全透明的,无法感受到此类服务和实际实时服务的差异。除了执行各种算法带来的计算开销外,也不会产生额外的信令开销或存储空间开销。

3 虚拟实时服务的关键模块及算法

本节具体地细化介绍了虚拟实时服务中各个功能模

项目简介

围绕具有分布天线、可重构、非集中控制特征的 environment-aware heterogeneous C-RAN 综合各类先进通信算法与协议。引入数值计算、受控博弈、组合优化等数学工具,实现立体网络的精准资源管理,尝试学科范式的转变。本项目拟重点突破如下科学任务,即超需求超密多层室内接入,通过突破关键性能参数的质变门限,向满足人类通信终极需求而努力,其指标参数如下:

- 超密集——高于室内典型人口密度(数量级 ~ 1 人/ m^3)
- 超需求——高于人眼分辨率高清视频(8K UHD TV, 100 Mbit/s);
- 若用户全部激活,并且可用带宽 100 Mbit/s,则所需的单位体积频谱效率约为 1~10 bit/s/MHz/ m^3 ,高于 5G 系统提出的相关性能指标。

以此科学任务明确具体参数设置和性能指标,理论联系实际,带动科学问题研究。希望重点突破“立体网络的精准资源管理”这一学科范式,从学术层面反思通信理论与工业应用的距离,提升对实际系统的描述精度和可应用性。

块的结构,并给出了示例算法;还从半定量的角度,对其性能特别是频谱效率和能量效率增益进行了评估和分析。

3.1 用户端的缓存内容保留选择算法

主张和倡导对终端的缓存内容选择算法建立工业标准,以便推送算法设计时有统一的标准。因此,该算法可以看作是系统实施的基准。图2给出了用户端的缓存区内容保留选择算法,其中外挂了内容需求时延的统计特性的学习模块,将在第3.3节重点介绍。如前文所述,新推送内容的到达、空闲缓存空间的减小,均可能导致用户存储单元的溢出。因此,用户终端需要评估用户对内容需求的程度,以便选择保留哪些内容,舍弃哪些内容。为此,提出了一种用户对内容的需求度评估方法,进而按需求度从高到低的

顺序保留内容,直到可用缓存空间不足时舍弃余下的内容。

接下来的问题就是:如何将用户对内容的需求度进行量化评估?提出如下思路:通过用户对内容的需求时延的概率分布,计算当前对该内容的需求度。此处,有两点需要考虑:从当前时刻算起,未来越有可能被用户提出使用请求的内容,其需求度应当越高;希望优先保留从当前时刻算起,“近期”需求的内容,避免某些很长时间以后才需要的内容对存储空间长期占用,并导致其他近期需要的内容因缺乏存储空间而被丢弃。基于上述两点考虑,定义用户 k 对内容 i 的有效需求度 d_{ik} 的计算式为:

$$d_{ik} = \int_0^{\infty} w(t) dF_{ik}(t + \tau_i) \quad (1)$$

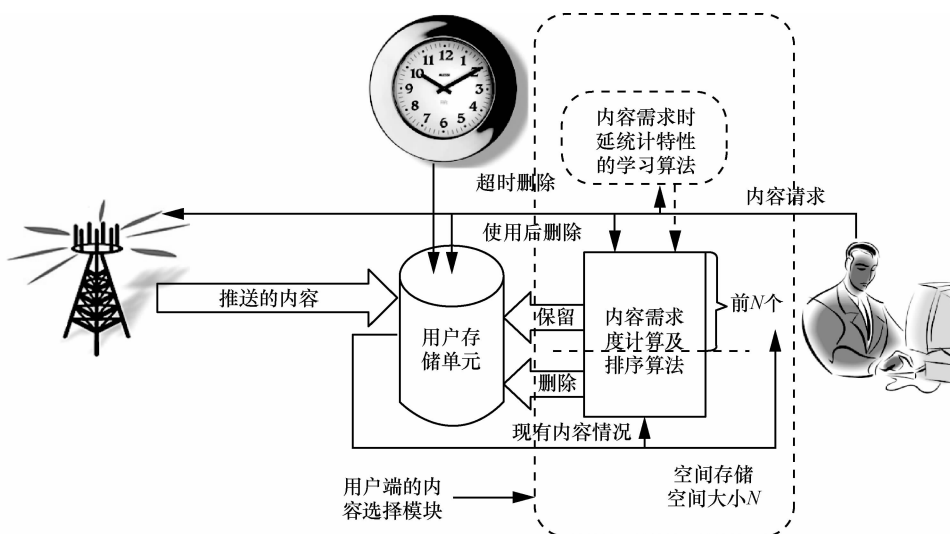


图2 用户端的缓存区内容保留选择模块

学术团队

国家“973”计划项目“密集立体覆盖移动通信的基础理论与方法”的研究任务由清华大学、北京邮电大学、东南大学共同承担。

依托的实验室包括:清华大学微波与数字通信国家重点实验室、清华大学智能技术与系统国家重点实验室、东南大学移动通信国家重点实验室、北京邮电大学泛网无线通信教育部重点实验室。

项目承研人员来自通信与信息系统领域(陈巍教授、彭木根教授、王东明副教授、王家恒副教授)和模式识别与智能系统领域(高飞飞副教授),能够有效实现跨学科交叉研究。



其中, τ_i 表示该内容已产生的时间; 加权函数 $w(t)$ 是一个满足 $\int_0^\infty w(t)dt$ 的非负减函数, 它表明更看重用户的近期需求。若 $w(t)$ 是一个从 0 到 T 的矩形窗, 则式(1)表示内容 i 从当前时刻起, 到时延 T 的未来时段内被用户 k 提出使用请求的概率。定义了用户 k 对内容 i 的需求度定量评估方法后, 用户终端可以对需要缓存的内容按照其需求度 d_i 的大小进行排序。若当前可用的存储单元缓存空间只能存放 I 个内容, 则选择保留需求度为 d_i 且排在前列的 M 个内容, 将其他需求度排序靠后的内容舍弃。

3.2 基站端的推送决策算法

基站端的内容推送决策模块如图 3 所示, 它包含两个主要的子模块: 基站端的内容需求度计算及排序算法; 推送发起的决策模块。其中, 还外挂了信道空闲和能量到达的统计特性的学习算法模块, 将在第 3.3 节重点介绍。

基站端的内容需求度计算及排序算法分两步完成。第一步是评估各个不同的用户对同一内容 i 的需求度, 其思想和方法类似于终端的内容需求度分析算法, 但是需要考虑两个特殊因素: 首先, 用户 k 终端中已有的内容或用户已经请求过的内容, 均无需重复进行推送, 因此, 若用户 k 已经请求过或已经存储了内容 i , 其需求度 $d'_{ik}=0$; 其次, 需

要考虑内容推送所需的传输时延 α 。假设推送传输一个内容需要 S 个时隙, 若采用固定划分的时隙及电网能量进行推送传输, 则 α 为恒定常数 S ; 若采用主系统的空闲时隙或者收割的可再生能源进行推送传输, 则 α 为一非负随机变量。为了定量刻画传输时延 α , 在一个简单情况下给出其概率分布函数。

假设主系统的空闲时隙或者收割的可再生能源在时间上独立, 服从一个伯努利过程, 记主系统时隙空闲的概率, 或者有一份能支撑本时隙传输的能量到达的概率为 θ , 此时 α 服从负二项分布, 写成连续时间的概率分布为:

$$g(\alpha) = \binom{U-1}{S-1} \theta^S (1-\theta)^{U-S} \delta(\alpha - U\Delta) \quad (2)$$

其中, Δ 为一个时隙的长度。若采用上述的半透明传输模式, 则从所需的 S 个时隙中扣除固定划分给虚拟实时服务的时隙数即可。此外, 若所需的时隙数 S 较大, 根据中心极限定理, 也可将推送传输的时延近似为一个恒定值 $\frac{S}{\theta}$ 。

由式(2)可以进一步预测用户 k 对内容 i 的平均需求度, 如下:

$$d'_{ik} = \int_0^\infty \int_0^\infty w(t+\alpha) g(\alpha) dF_i(t+\tau_i+\alpha) \quad (3)$$

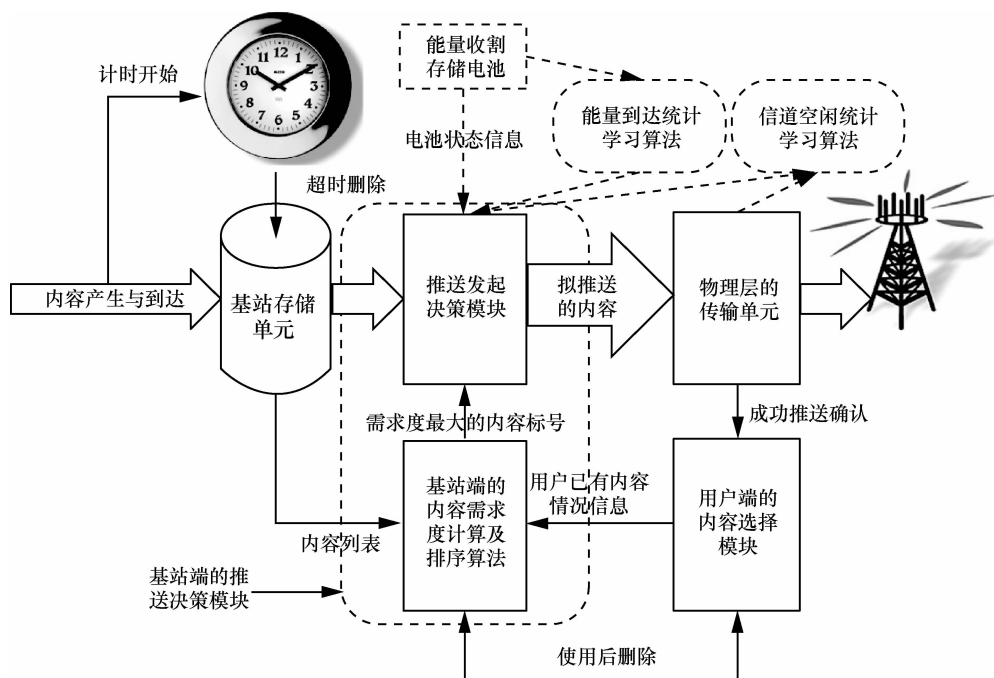


图3 基站端的推送决策模块

在完成了预测用户 k 对于内容 i 的平均需求度之后,内容需求度计算及排序算法进入其第二步。注意到即使对于相同的内容 i ,不同的用户由于其时延分布函数、内容请求历史等不同,具有不同的预测需求度 d'_{ik} 。如何根据需求度矩阵 $d'=[d'_{ik}]$,综合考虑用户缓存区的剩余空闲空间,选出从时间集合和用户集合双重角度来看,平均需求度最高的内容 i^* ,就是接下来要讨论的问题。为此,定义内容 i 的加权总需求度为:

$$\bar{d}'_i = \sum_{k=1}^K \omega(I_{kn}) d'_{ik} \quad (4)$$

其中, I_{kn} 表示用户 k 在当前时隙为 n 时,存储区中的空闲空间。 $\omega(I_{kn})$ 是一个加权系数函数,值域为 $[0, 1]$, 且是 I_{kn} 的增函数。这样设定的目的是,倾向于给尚有较大存储空间的用户推送内容,尽可能避免推送的内容因缓存空间不足而被丢弃。考虑一个特殊情况,若用户 k 的缓存区已满,即 $I_{kn}=0$,且 d'_{ik} 小于其缓存区中已有内容的需求度(或者其考虑随机推送时延后的预测值),则即使向该用户推送了内容 i ,也会以很大概率被丢弃,造成传输资源的浪费,此时应当设置 $\omega(I_{kn})=0$ 。

基于式(4)计算得到的内容 i 的加权总需求度 $\bar{d}'_i$,可知若发起推送,则基站最应当推送的内容为:

$$i^* = \arg \max_i \bar{d}'_i \quad (5)$$

推送发起决策模块所做的决策是——是否推送这个最应当推送的内容 i^* 。本文提出一种简单的门限判别策略,即设定一个推送决策门限 D ,若 $\bar{d}'_{i^*} > D$,则推送内容 i^* ;否则,基站当前不发起任何推送。

在门限判别策略中,阈值 D 越大,虚拟实时服务的推送越保守。此时,虽然减少了对空闲频谱的使用,但是也避免了推送用户不迫切需求的内容,或者推送内容过多导致缓存区溢出的情况,因此可以减少缓存区溢出丢弃内容而造成的能量浪费,其能量效率也得到了提高;与此相反,阈值 D 越小,虚拟实时服务的推送越激进。此时,基站抓住尽可能多的空闲时隙推送内容,但这样会降低内容的命中率,推送出一些用户需求度偏低的内容,并且因为推送量较大而更多地出现用户缓存区溢出、内容被丢弃的情况。因此,虽然减小阈值 D 提升了频谱利用率,但是其能量效率也相应降低了。

由上述定性分析可以看出,在虚拟实时服务体制中,频谱效率和能量效率之间存在折中关系。当然,对于采用能量

收割通信的多播推送,由于收割来的可再生能源是“免费”的,若不加利用反而是浪费。因此,阈值 D 的设计需要考虑电池的状态进行动态调整。当电池能量较多时,应降低门限阈值 D 以进行更积极的推送;当电池能量较少甚至被消耗一空的时候,应提高门限阈值 D 以进行更为保守的推送。

从上面的讨论中还可以发现,当各个用户的剩余存储空间总体上来说比较多时,基站会更积极地推送更多的内容,反之则减少推送的内容,这一点有些类似于 TCP 中的流量控制。此外,在内容的生命周期尚未结束时,被推送过的内容还需保留在基站的存储单元中。这是因为不同用户对同一内容的需求时延的概率分布函数不同,有可能出现这种情况——对某一内容的需求度在不同时刻多次出现峰值,使得该内容不止一次地被基站选择推送。因此,有必要在基站缓存区中保留已经被推送过的业务。

3.3 各类统计特性的实时学习算法

在虚拟实时服务协议的实现中,涉及 3 个外挂的统计特性的学习模块,分别是:用户对内容的需求时延统计特性的学习模块、终端空闲存储空间的统计特性的学习模块、信道空闲和能量到达的统计特性的学习模块。其中,对于信道空闲和收割能量到达的统计特性的学习策略,在对认知无线电技术和能量收割通信技术的研究中,已经取得了较多成果^[26,27],这里不再赘述。

首先,介绍用户对内容的需求时延的统计特性学习算法。上文中,一个基本的前提假设是,用户 k 对内容 i 的需求时延的概率累积函数 $F_{ik}(x) = \Pr\{T_{ik} \leq x\}$ 是已知的。满足这一分布的内容 i 可以聚合为一类,并且使用某种内容标签进行标识。在此基础上,学习模块可以记录用户 k 对属于该类内容,并且排序在第 l 条产生的请求时延 t_l ,初始化 $F_{ik}(x)=0$,则概率累积函数 $F_{ik}(x)$ 可估算为:

$$F_{ik}(x) = \frac{1}{l} F_{ik}(x) + \frac{1}{l} H(x - t_l) \quad (6)$$

其中, $H(x)$ 表示单位阶跃函数。

用户空闲存储空间的统计特性取决于用户 k 对自身设备的使用习惯。一般来说,剩余空间的大小可以用一个 Markov 链进行描述,其更新方程为:

$$I_{kn} = I_k(n-1) - Q_{kn} \quad (7)$$

其中, Q_{kn} 为用户 k 在第 n 个时隙的存储量,取正值表示用户向存储单元中存入其他内容,取负值表示用户从存



储单元中删除其他内容。 Q_{kn} 的概率分布函数记为 $G_{kn}(x) = \Pr\{Q_{kn} \leq x\}$, 可估算为:

$$G_{kn}(x) = \frac{1}{n} G_{k(n-1)}(x) + \frac{1}{n} H(x - Q_{kn}) \quad (8)$$

当引入在线的需求估计时,系统的内部状态估计和运行控制相互作用,形成了类似 Kalman 滤波的更新迭代模式,在理论上具有较大的研究空间。由此可见,虚拟实时服务产生了很多新的理论问题,将在最后一节简要介绍可能突破的理论问题。

最后,考虑一种特殊情况,即 $F_{ik}(x) = H(x)$,以验证虚拟实时服务的合理性。显然, $F_{ik}(x) = H(x)$ 的业务就是即时通信业务,如语音电话、视频电话、现场直播等,代入上述的算法和协议构架中可以看到:当业务产生时,就应当立刻发起传输,而由于即时通信的内容在产生后不久就会失去价值,到达用户后也无需进行缓存。可见,虚拟实时服务可以在 $F_{ik}(x) = H(x)$ 的情况下退化为即时通信。

4 结束语

提出了一种基于虚拟实时服务的协议构架,对其关键模块与算法进行了综述和分析。为了分析用户对内容的需求度,提出了一种基于用户对内容需求时延的统计分布,量化描述用户需求的方法。在此基础上,提出了用户对缓存内容需求度的定量分析和排序的方法,由此需求度排序实现缓存区的内容保留选择算法。类似于增量调制等信源编码的构架,建议对用户端的缓存内容保留选择算法进行标准化,在此基础上设计基站端的推送决策算法,可利用标准化的用户端的缓存内容保留选择算法在基站形成本地预测反馈,从而避免因推送过多而产生大量内容丢弃或不必要的重复推送等情况,选择实际需求度最大的内容进行推送。对上述算法进行了简要的量化描述,半定量分析了其中存在的能量和频谱效率折中,并发掘了基站推送决策与 TCP 的类似之处。此外,还对系统中辅助的各类统计特性的实时学习算法进行了介绍。

在虚拟实时服务的演进过程中,尚有如下理论问题待解决。首先,用户如何学习其对内容的需求时延的统计特性,是一个较困难的问题。第 3.3 节中提出的在线学习算法有一个基本前提——已经明确知晓内容的分类,即同一类的内容对于用户 k 具有相同的统计特性,然而,如何对内容进行标签和分类是一个更为困难的问题,涉及模式识

别、人工智能以及内容(包括视频、语音等)的提取与理解等多个方面;其次,由于基站端的推送内容决策模块存在反馈回路,统计特性的学习模块也会引起控制反馈,因此,系统的反馈控制稳定性有待阐明,特别是实时估算用户的内容需求时延分布函数时,系统内部存在多个带有误差的反馈回路,其稳定性条件的理论分析具有广阔的研究空间;第三,内容需求、信道空闲和可再生能源的时间非平稳特性对系统的运行会带来影响。一般来说,凌晨时刻内容需求的时延较大,信道空闲的概率较大;而太阳能等可再生能源在白天较为充足,因此在较大的时间尺度上来看,它们的统计特性并非时间平稳的;第四,由于用户的信道条件、存储空间等在实际中存在异构特性,公平性保障方法需要进一步研究。本文提出的算法,主要以提升能量和频谱效率为目标,并没有考虑公平性。而对于存在异构特性的情况,不应当简单地对各个用户的需求度采用相同的加权函数进行求和,而是应当考虑用户的服务历史以及缓存区大小不同的统计特性,对信道条件较差、缓存区相对较小、内容需求度与大众差异较大的用户给予适当倾斜;第五,物理层的时变特性,以及由此提供的自适应调制编码等传输自由度,可以与虚拟实时服务进行跨层联合优化;最后,本文所讨论的虚拟实时服务主要集中于协议构架和算法层面,这一全新服务体制的诸多基础理论性能分析问题,如频谱和能量效率极限等问题还有待进一步探索。

参考文献

- 1 Andrews J G, Buzzi S, Choi W, *et al.* What will 5G be. *IEEE Journal on Selected Areas in Communications*, 2014, 32 (6): 1065~1082
- 2 Hanzo L, Haas H, Imre S, *et al.* Wireless myths, realities, and futures: from 3G/4G to optical and quantum wireless. *Proceedings of the IEEE Special Centennial Issue*, 2012: 1853~1888
- 3 Kedar D, Arnon S. Urban optical wireless communication networks: the main challenges and possible solutions. *Communications Magazine*, 2004, 42(5): S2~S7
- 4 Richardson T J, Shokrollahi M A, Urbanke R L. Design of capacity-approaching irregular low-density parity-check codes. *IEEE Transactions on Information Theory*, 2001, 47 (2): 619~637
- 5 Cimini L J. Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing. *IEEE Transactions on Communications*, 1985, 33(7): 665~675

- 6 Liu J, Chen W, Zhang Y J, *et al.* A utility maximization framework for fair and efficient multicasting in multicarrier wireless cellular networks. *IEEE/ACM Transactions on Networking*, 2013, 21(1): 110~120
- 7 Foschini G J. Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. *Bell Labs Technical Journal*, 1996, 1(2): 41~59
- 8 Foschini G J, Gans M J. On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Personal Communications*, 1998, 6(3): 311~335
- 9 Gutierrez F, Agarwal S, Parrish K, *et al.* On-chip integrated antenna structures in CMOS for 60 GHz WPAN systems. *IEEE Journal on Selected Areas in Communications*, 2009, 27 (8): 1367~1378
- 10 Rappaport T S, Sun S, Mayzus R, *et al.* Millimeter wave mobile communications for 5G cellular: It will work. *Access IEEE*, 2013(1): 335~349
- 11 Cadambe V R, Jafar S A. Interference alignment and degrees of freedom of the-user interference channel. *IEEE Transactions on Information Theory*, 2008, 54(8): 3425~3441
- 12 Andrews M, Mitra P. Tripling the capacity of wireless communications using electromagnetic polarization. *Nature*, 2001, 409(6818): 316~318
- 13 陈巍, 王东明, 王家恒. 密集立体覆盖: 未来移动通信的机遇和挑战. *电信科学*, 2013, 29(6): 2~9
Chen W, Wang D M, Wang J H. 3-dimensional dense coverage: opportunities and challenges for future mobile communications. *Telecommunications Science*, 2013, 29(6): 2~9
- 14 Wang D, Chen W, Wang J, *et al.* 3-dimension coverage with ultra-densely distributed antenna systems: system design, rate analysis. *Proceedings of IEEE International Conference on Communications (ICC)*, London, UK, 2015
- 15 Yan S, Peng M, Chen W, *et al.* Downlink heterogeneous small cell networks with cell associations in k-floor indoor scenarios. *Proceedings of IEEE International Conference on Communications (ICC)*, London, UK, 2015
- 16 Haykin S. Cognitive radio: brain-empowered wireless communications. *IEEE Journal on Selected Areas in Communications*, 2005, 23(2): 201~220
- 17 Ulukus S, Yener A, Erkip E, *et al.* Energy harvesting wireless communications: a review of recent advances. *IEEE Journal on Selected Areas in Communications*, 2015, 33(3): 360~381
- 18 Liu J, Chen W, Cao Z, *et al.* Cooperative beamforming for cognitive radio networks: a cross-layer design. *IEEE Transactions on Communications*, 2012, 60(5): 1420~1431
- 19 Liu J, Dai H, Chen W. Delay optimal scheduling for energy harvesting based communications. *IEEE Journal on Selected Areas in Communications*, 2015, 33(3): 452~466
- 20 Zhou S, Cheng T, Chen W. Outage minimization for a fading wireless link with energy harvesting transmitter, receiver. *IEEE Journal on Selected Areas in Communications*, 2015, 33(3): 496~511
- 21 Liu J, Chen W, Cao Z, *et al.* Dynamic power and sub-carrier allocation for OFDMA-based wireless multicast systems. *Proceedings of IEEE International Conference on Communications*, Beijing, China, 2008: 2607~2611
- 22 陈巍. 非时齐多元协同通信: 虚拟实时服务理论及应用. *电信科学*, 2014, 30 (7): 43~52
Chen W. Non-homogeneous multi-Unit collaborative communications: a theory and application for virtual real-time service. *Telecommunications Science*, 2014, 30(7): 43~52
- 23 Zhou S, Gong J, Zhou Z, *et al.* Greendelivery: proactive content caching and push with energy harvesting based small cells. *IEEE Communications Magazine*, 2015, 53(4): 142~149
- 24 Wang K, Chen Z, Liu H. Push-based wireless converged networks for massive multimedia content delivery. *IEEE Transactions on Wireless Communications*, 2014, 13(5): 2894~2905
- 25 Ellison N B. Social network sites: definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 2007, 13(1): 210~230
- 26 Mitola J, Maguire G Q. Cognitive radio: making software radios more personal. *Personal Communications*, 1999, 6(4): 13~18
- 27 Kansal A, Hsu J, Zahedi S, *et al.* Power management in energy harvesting sensor networks. *ACM Transactions on Embedded Computing Systems (TECS)*, 2007, 6(4): 1~35

[作者简介]



陈巍,男,博士,清华大学电子工程系教授、博士生导师(工学、工程)、副系主任、教学委员会主任、清华大学“电子与通信工程”学位评定分委员会副主席,中国通信学会青年工作委员会副主任委员,中国电子学会教育工作委员会青年工作组副主任委员,北京市通信学会青年工作委员会委员,IEEE Transactions on Education、IEEE Wireless Communication Letters、China Communications 等国际期刊编委,“全国五一劳动奖章”、“北京市优秀青年人才”、“北京市优秀教师”荣誉称号获得者,国家重点基础研究发展计划(“973”计划)首席科学家,国家自然科学基金优秀青年科学基金、第十四届霍英东高校青年教师奖一等奖、第十七届茅以升北京青年科技奖获得者,IEEE 高级会员,国家中青年科技创新领军人才计划、教育部新世纪优秀人才支持计划、北京市科技新星计划和清华大学“221”人才计划入选者。主要研究方向为无线通信、信息论、移动互联网。