算力网络中面向业务体验的算力建模

Computing Power Modeling for Business Experience in Computing Power Network

李建飞 /LI Jianfei，曹畅 /CAO Chang，李奥 /LI Ao，庞博文 /PANG Bowen
（中国联通研究院，中国 北京 100048）
(China Unicom Research Institute, Beijing 100048, China)

摘要：针对算力网络中算力量化的问题，对异构的 IT 算力资源进行归一化建模，并提出算力的分级标准。同时阐述了为保障业务体验的算力、存储、网络等的联合服务能力，从业务角度归纳了不同类型业务的服务能力需求，旨在形成通用的算力服务，为客户的业务体验提供基础保障。

关键词：算力网络；算力建模；算力分级；业务需求

Abstract: In order to solve the problem of computing power in the computing power network, a normalized model of heterogeneous IT computing power resources is established, and a classification standard for computing power is proposed. The joint service capabilities of computing power, storage and network to ensure business experience are discussed, and the service capability requirements of different types of business from the business perspective are introduced, aiming to form a general computing power service and provide basic guarantee for customers' business experience.

Keywords: computing power network; computing power modeling; computing power hierarchy; business requirements

DOI：10.12142/ZTETJ.202005007
网络出版地址：https://kns.cnki.net/kcms/detail/34.1228.TN.20200927.1351.002.html

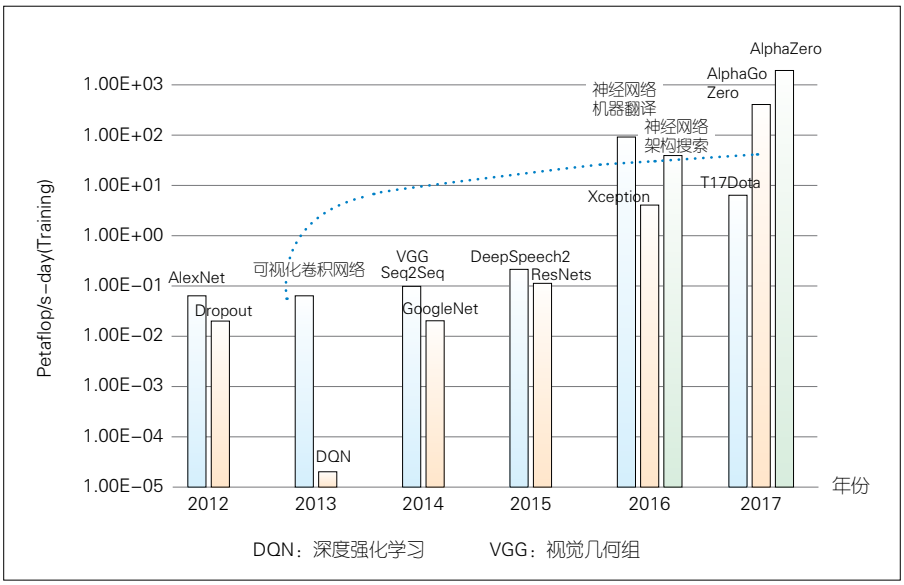
网络出版日期：2020-09-27
收稿日期：2020-08-20

人工智能（AI）是一项引领未来的技术。近年来，随着深度学习、大数据、群体智能等技术在智慧医疗、智慧教育、智能安防、智能制造、智能巡检等领域的广泛应用，人工智能已经成为当代社会一项通用的技术。算法、数据和算力共同组成人工智能的三要素。一直以来，算力以不同的形式存在于人类发展的各个阶段，从古代的算盘到机械式计算器、电子计算器，再到晶体管、移动电话[1]，算力已经渗透到人们生活中的方方面面。

基金项目：国家重点研发计划（2019YFB1802800、2019YFB1802600）

算力既是 AI 的基础，也是 AI 发展的主要驱动力。如同驱动前两次工业革命的煤炭和电力一样，算力也驱动着人工智能的革命不断前行。在 20 世纪 70 年代，虽然人工神经网络模型的理论架构已经基本成熟，却在之后的几十年里一直没能得到认可和应用，直到近来才得以“重见天日”，其中的根因就在于算力的限制，即当时的算力无法有效支撑算法的运行。在算法和数据确定的情况下，算力的增加可以使算法获得更好的训练效果，同时大大减少有效的训练时间。据统计（如图 1 所示），自 2012 年以来人们对于算力的需求增长超过 30 万倍（而如果按照摩尔定律的速度，只有 12 倍的增长）。

在算力网络时代，网络与算力相融合作为基础资源提供服务。运营商基于算力网络[2-4]，为客户提供所需算力和确定时延的产品。网络为计算服务的价值在于释放算力。目前，各种已经兴起的（例如虚拟现实/增强现实）和潜在的（例如自动驾驶）智能业务，均对算力提出了较高的要求，但是针对信息技术（IT）基础设施，其面向业务所提供的算力需求并没有量化，也没有针对算力需求的分级。本文中，我们对异构的 IT 算力资源进行归一化建模，并且提供算力的分级标准，以



▲图 1 从 AlexNet 到 AlphaGo Zero 训练类算力需求增长 30 万倍

便算力提供者在设计业务套餐时进行参考。

1 算力网络中算力的衡量与模型

算力的衡量与建模是提供算力服务的基础。将底层异构算力资源量化建模，能够形成业务层可理解、可快速使用的统一量化的算力资源。

1.1 算力定义

算力是近年来业界讨论的热门话题，但对“算力是什么”这个问题一直没有一个通用标准的定义。2018 年，诺贝尔奖获得者、经济学者 WILLIAM D. N. 在《计算过程》中对算力进行定义：算力是设备根据内部状态的改变，每秒可处理的信息数据量。

本文中，算力被定义为：算力是设备或平台为完成某种业务所具备的处理业务信息的关键核心能力。它涉及设备或平台的计算能力，包括逻辑运算能力、并行计算能力、神经网络加速能力等。

1.2 算力衡量指标

根据所运行算法和涉及的数据计算类型，算力可被分为逻辑运算能力、并行计算能力和神经网络计算能力。

（1）逻辑运算能力。

这种计算能力是一种通用的基础运算能力。硬件芯片代表是中央处理器（CPU），这类芯片需要大量的空间去放置存储单元和控制单元。相比之下，计算单元只占据了很小的一部分。因此，它在大规模并行计算能力上很受限制，但可以用于逻辑控制。一般情况下，TOPS（表示处理器每秒钟可进行一万亿次操作）被用来衡量运算能力。在某些情况下，能效比 TOPS/W（表示在 1 W 功耗的情况下，处理器能进行多少次操作）也可被作为评价处理器运算能力的一个性能指标。

（2）并行计算能力。

并行计算能力是指专门为了处理如图形图像等数据类型统一的一种高效计算能力，是一种比较通用的计算能力。这种计算能力特别适合处理大量的类型统一的数据，不仅在图形图像处理领域大显身手，同时还适用于科学计算、密码破解、数值分析、海量数据处理（排序、Map-Reduce 等）、金融分析等领域。

典型的硬件芯片代表是英伟达推崇的图形处理单元（GPU）。GPU 的构成相对简单，有数量众多的计算单元和超长的流水线。浮点运算能力常被作为并行计算的度量标准。单位 TFLOPS/s 可以简单写为 T/s，意思是一万亿次浮点指令每秒。此外，相关单位还有 MFLOPS、GFLOPS、PFLOPS。

（3）神经网络计算能力。

神经网络计算能力主要用于 AI 神经网络、机器学习类密集计算型业务，是一种用来对机器学习、神经网络等进行加速的计算能力。

近年来，厂商发布的 AI 类芯片都是为加速神经网络计算而设计的，例如华为技术有限公司的网络处理器（NPU）、Google 公司的张量处理单元（TPU）[5]。另外，机器学习、神经网络的本质是密集计算。Google 公司工程师认为：如果人们每天用 3 min 的语音搜索，但在运行时没有 TPU 加持的语音识别人物的话，运营公司将需要建造两倍多的数据中心。

专门做神经网络计算能力的芯片厂商都有各自测试的 Benchmark，处理能力也大多是配合各自研发的算法。目前，这类能力常用的度量单位也是浮点计算能力 FLOPS。浮点运算能力高的计算设备能够更好地满足在同一时间里更多用户的任务需求，可以更有效地处理高并发任务数量的业务。

1.3 算力量化模型

算力的统一量化是算力调度和使用的基础。如前所述，算力的需求可分为 3 类：逻辑运算能力、并行计算能力以及神经网络加速能力。同时对不同的计算类型，不同厂商的芯片也各自不同的设计，这就涉及异构算力的统一度量[6-7]。不同芯片所提供的算力可通过度量函数映射到统一的量纲。

针对异构算力的设备和平台，假设存在 n 个逻辑运算芯片、m 个并行计算芯片和 p 个神经网络加速芯片，那么业务的算力需求如公式（1）所示：

$$C_{\mathrm{br}}=\begin{cases}\sum_{i=1}^{n}\alpha_i\cdot f(a_i)+q_1(\mathrm{TOPS}) & \text{逻辑运算能力}\\ \sum_{j=1}^{m}\beta_j\cdot f(b_j)+q_2(\mathrm{FLOPS}) & \text{并行计算能力}\\ \sum_{k=1}^{p}\gamma_k\cdot f(c_k)+q_3(\mathrm{FLOPS}) & \text{神经网络加速能力}\end{cases}\quad(1)$$

公式（1）中，C_{br} 为总的算力需求，$f(x)$ 是映射函数，α、β、γ 为映射比例系数，q 为冗余算力。以并行计算能力为例，假设有 b_1、b_2、b_3 3 种不同类型的并行计算芯片资源，则 $f(b_j)$ 表示第 j 个并行计算芯片 b 可提供的并行计算能力的映射函数，q_2 表示并行计算的冗余算力。

2 算力网络中算力分级

随着 AI、5G 的兴起，各种智能业务也应运而生，并呈现多样化趋势[8]。不同的业务运行所需的算力需求的类型和量级也不尽相同，例如非实时、非移动的 AI 训练类业务。这类业务训练数据庞大，神经网络算法层数复杂，若想快速达到训练效果，需要计算能力和存储能力都极高的运行平台或设备。对于实时类的推理业务，一般要求网络具有低时延，但对计算能力的需求则可降低几个量级。将业务运行所需的算力按照一定标准划分为多个等级，不仅可供算力提供者在设计业务套餐时参考使用，还可以为算力平台设计者在设计算力网络平台时提供算力资源选型依据。

由于智能应用对算力的诉求主要是浮点运算能力，因此，业务所需的浮点计算能力的大小可作为算力分级的依据。针对目前应用的算力需求，可将算力划分为 4 个等级，具体如表 1 所示。

从现有业务上看，超算类应用、大型渲染类业务对算力的需求是最高的，可达到 P 级的算力需求，这类需求被定位为超大型算力；大型算力主要是 AI 训练类应用，根据算法的不同以及训练数据的类型和大小，这类应用所需的算力从 T 级到 P 级不等；小型算力则主要是针对类似 AI 推理类业务，这类业务大多部署在终端边缘，对算力的需求稍弱，从几百 G 到 T 级不等；此外，小于 500 GFLOPS 的算力需求被定义为小型算力。

3 面向业务体验的算力、存储、网络等联合服务

业务运行需要平台或设备的算力需求保障，同时不同类型的业务还需要诸如存储能力、网络服务等个性化能力[8]。

3.1 面向业务体验的联合服务能力

（1）存储能力。在算力网络中，存储在数据处理过程中起到至关重要的作用。随着数据处理需求的日益增长，数据存储的重要性也显著提升。内存与显存的数量可以作为关键指标被用来衡量计算存储的能力，通常以吉比特为单位。存储能力在很大程度上会影响计算机的处理速率。

（2）网络能力。在保障业务服务质量（QoS）方面，网络性能是一个非常重要的指标（尤其是针对一些实时性业务），这就需要灵活调度部署网络以满足业务对时延和抖动的需求。对于人工智能应用来说，模型的推理时延也是衡量算力的关键指标。推理时延越低，用户的体验越好，而较高的时延可能会导致某些实时应用无法达到要求。

（3）编解码能力。编解码能力是利用设备或者程序对信号或数据流进行变换的能力。这里的变换既包括将信号或者数据流进行编码或提取得到编码流的操作，也包括为了观察或者处理而进行的其他操作。编解码器经常用在视频会议和流媒体等涉及图形图像处理的应用中。

编解码相应的硬件需要编码解码的引擎配置。一般的编解码能力附着在计算芯片上，如英伟达 GPU 芯片带有编解码引擎（编码引擎为 NVENC，解码引擎为 NVDEC）。

（4）每秒传输帧数（FPS）。FPS 主要用于渲染场景，属于图像领域的定义，它是指画面每秒传输的帧数，即动画或者视频的画面数。每秒能够处理的帧数越多，画面就会越流畅。在分辨率不变的情况下，GPU 的处理能力越高，FPS 就越高。

（5）吞吐量。在深度学习模型的训练过程当中，一个关键指标就是模型每秒能输入和输出的数据量。在广大的 AI 应用中，图像和视频业务占据了很高的比例，因此，在衡量吞吐量的时候，我们可以使用 Images/s 这个单位来衡量模型的处理速度。

设备或平台的运行业务的服务能力涉及前文所述的算力、网络和存储，

▼表 1 算力分级表

算力分类等级	算力水平	典型推理场景
超大型算力	>1 PFLOPS，P 级算力	渲染农场、超算类应用；部分大型模型训练，如 VGGNet 模型训练
大型算力	10 TFLOPS~1 PFLOPS	多数模型训练，如 CNN、RNN 训练
中型算力	500 GFLOPS~10 TFLOPS	推理类应用，如安防、目标检测
小型算力	< 500 GFLOPS	小型计算应用场景、单条语音语义

注：1 GFLOPS=10^9 FLOPS，1 TFLOPS=10^{12} FLOPS，1 PFLOPS=10^{15} FLOPS

CNN：卷积神经网络　RNN：递归神经网络　VGG：视觉几何组

以及其他能力（如 FPS、吞吐量等）。这些能力共同保障着用户的业务体验。

3.2 不同业务场景的服务能力需求

（1）训练类场景。

训练业务是指通过大数据训练出一个复杂的神经网络模型，即用大量标记过的数据来“训练”相应的系统，使之适应特定的功能场景。训练不仅需要极高的计算性能，还需要处理海量数据，同时也要具有一定的通用性，以便完成各种各样的学习任务。目前训练业务主要集中在云端，需要有足够强的计算能力[9-10]作为保障。训练类业务的服务能力需求如表 2 所示。

（2）推理类场景。

推理类业务是指利用训练好的模型，使用新数据推理出各种结论，即借助现有神经网络模型进行运算，利用新的输入数据一次性获得正确结论的过程，也叫作预测或推断。虽然目前推理过程主要在云端完成，但越来越多的厂商正将其逐渐转移到终端[9]。推理对计算性能要求不高，但更注重综合指标，如单位能耗算力、时延、成本等。推理类业务的服务能力需求如表 3 所示。

（3）云增强现实（AR）/ 虚拟现实（VR）类场景。

移动 AR/VR 业务是一种云、端相结合的方式，其本质是一种交互式在线视频流[11]。对于云侧拥有超强算力和低延时的网络，更多的渲染工作首先在云侧完成，然后再通过网络传送给用户侧，如手机、PC、PAD、机顶盒等终端设备。用户通过输入设备（虚拟键盘、手柄等）对业务进行实时操作，如图 2 所示。

另外，在高铁、地铁等高速移动的场景下，用户侧终端设备将会在多个基站甚至多个地域进行网络切换，这将导致初始连接的云侧节点网络延迟增加。根据用户的实际情况进行统一的调度和管理，将计算能力在多个节点之间无缝迁移，可保障流畅切换的无感用户体验。此外，爆款的 AR/VR 游戏通常会在短时间内汇聚大量用户，其社交属性会带来地域相对密集的特点，这就要求算力网络节点能够快速调用计算能力、设计灵活架构、实现弹性伸缩，以满足用户的密集需求。云 VR/AR 业务的服务能力需求如表 4 所示。

（4）视频类场景。

伴随宽带网络和移动互联技术的不断提升，娱乐视频、通信视频、行业视频等各大领域的视频业务迅猛发

▼表 2 人工智能模型训练业务相关参数

具体算法	应用场景描述	算力需求估算 / PFLOPS（FP64）	网络需求估算	存储需求估算	备注
VGGNet	在数据集上训练网络模型，提升检测效果，以训练迭代一次为例	19	非实时类业务	VGG16 模型权重大小 138.37 MB；ImageNet 数据集大小 ~1 TB	VGG16 在 ImageNet 数据集为例
VGGNet	在数据集上训练网络模型，提升检测效果，以训练迭代一次为例	6	非实时类业务	VGG16 模型权重大小 138.37 MB；COCO 数据集大小 ~20 GB	VGG16 在 COCO 数据集为例
ResNet50	在数据集上训练网络模型，提升检测效果，以训练迭代一次为例	5	非实时类业务	ResNet50 模型权重大小 25.56 MB；ImageNet 数据集大小 ~1 TB	ResNet50 在 ImageNet 数据集为例
ResNet50	在数据集上训练网络模型，提升检测效果，以训练迭代一次为例	2	非实时类业务	ResNet50 模型权重大小 25.56 MB；COCO 数据集大小 ~20 GB	ResNet50 在 COCO 数据集为例

COCO：微软公司开发的一个数据集　　FP：浮点数精度　　VGG：视觉几何组

▼表 3 人工智能推理预测业务相关参数

具体算法	应用场景描述	算力需求估算	网络需求估算（时延）/ms	存储需求估算	备注
CNN	单张图像的人脸识别任务	10 GFLOPS（FP64）	/	MTCNN 模型权重大小：186 MB	MTCNN 人脸识别算法为例（CNN 约占 80% 算力需求）
	单路单流对人脸图像进行识别；应用在实验室环境	13 GFLOPS（FP64）	<60		
	单路多流对人脸图像进行识别；应用在写字楼等场景，实现并发（300 张图片并发为例）人脸识别功能	4 TFLOPS（FP64）	<60		
	多路多流对人脸图像进行识别（16 路，300 张图像并发）；应用在城市街道、闹市区	64 TFLOPS（FP64）	<200		
RNN	对一条语音进行语音识别	2 GFLOPS（FP64）	<60	DeepSpeech2 普通话语音识别模型权重大小：216 MB	DeepSpeech2 语音识别算法为例
	实现并发语音识别任务（以 500 条语音识别为例）	1 TFLOPS（FP64）	<60		

CNN：卷积神经网络　　FP：浮点数精度　　MTCNN：多任务卷积神经网络　　RNN：递归神经网络

展。除了传统的视频会议之外，视频培训、视频客户服务、远程医疗、在线直播等一系列新兴视频应用正在各个行业迅速普及[12]。视频类业务的服务能力需求如表4所示。

（5）智能驾驶场景。

智能驾驶、车联网是智慧城市的重要组成部分。在2019年新冠疫情出现时，无人车送餐、无人车消杀等都体现了智能驾驶的优势。考虑到智能驾驶对安全要求极高的特殊性[13]，目前每个车辆都装备有大算力的工控机，这大大增加了无人驾驶车辆的成本。若将车辆的计算能力释放到云侧，则需要算力网络同时具备极低的时延和超强的算力。此外，自动驾驶具有移动性，需要算力节点的无缝切换，以保障自动驾驶业务的超低时延。智能驾驶业务的服务能力需求如表4所示。

4 结束语

本文中，我们针对不同算力资源进行统一建模，给出了算力分级的标准，并阐述了为保障业务体验的算力、存储、网络等的联合服务能力，同时从业务的角度归纳了不同类型业务的服务能力需求。算力的衡量与建模是一个比较困难但却很重要的研究课题。在未来，随着算力（特别是边缘算力）的进一步扩大，算力与网络的结合将越来越紧密。通过网络对算力进行调度，引入合理的网络调度方法，可降低云边端协同的智能业务对算法和算力的需求。

参考文献

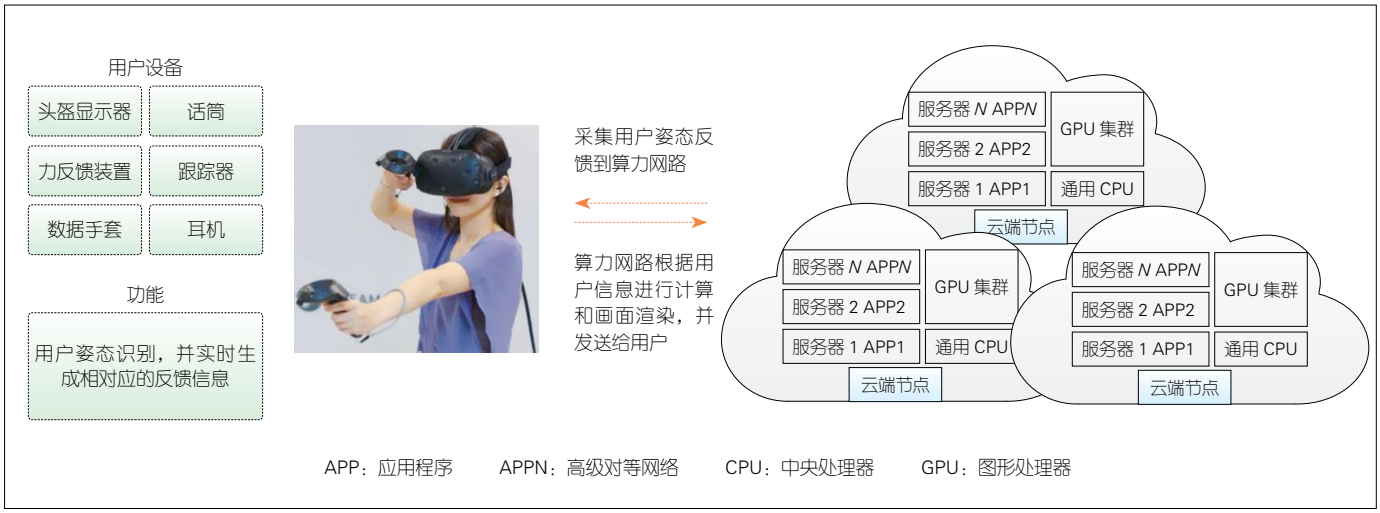
[1] 华为技术有限公司 . 泛在算力：智能社会的基石 [R/OL]. [2020-08-07]. https://www.huawei.com/cn/public-policy/ubiquitous-computing-power

[2] 中国联合网络通信有限公司 . 中国联通算力网络白皮书 [R/OL]. [2020-08-07]. http://www.bomeimedia.com/China-unicom/index-01.1.html

[3] 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

[4] 姚惠娟，耿亮 . 面向计算网络融合的下一代网络架构 [J]. 电信科学，2019, 35(9): 38-43

下转第52页 ➡



▲图2 虚拟现实系统组成及交互示意图

▼表4 云VR/AR、视频类和智能驾驶业务的服务能力需求

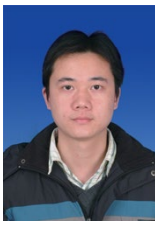
业务类型	应用场景描述	算力需求估算	网络需求估算	存储需求估算	备注
云AR/VR	PC VR、移动VR、2D AR动作本地闭环、全景云端下载、远程办公、购物等 VR新零售	40 EFLOPS（FP32）算力需求来自视频编解码以及视频内容语义感知和环境感知	20 Mbit/s 时延<50 ms 40 Mbit/s 时延<20 ms	运行环境： 内存4 GB，存储32 GB	算力依据具体场景而定
视频类任务	4 K、8 K点播业务 网红直播 视频会议双流	教育点播、娱乐点播等 直播业务员 实时视频业务	小型算力<500 GFLOPS	<15 ms，200 Mbit/s <150 ms，10 Mbit/s <25 ms，20 Mbit/s	运行环境：单个流媒体服务器（内存16 GB，存储空间~1 TB）
智能驾驶	环境感知 决策避障 自车定位	融合多路视觉激光等数据、推理计算 对障碍物轨迹跟踪、风险提醒 根据感知等信息给出自车6DOF位姿	24 TOPs/8 TFLOPS（Drive PX2）FP16	<5 ms <10 ms <5 ms	运行环境： 内存16 GB， 存储空间128 GB

6DOF：六自由度系统　AR：增强现实　FP：浮点数精度　PC：个人电脑　VR：虚拟现实

对其进行求解。在此基础上，设计了一个基于 AC 的基站功率动态智能控制算法。仿真实验结果证明，该算法能够有效降低超密集蜂窝网络中基站间的相互干扰，提升网络传输性能。

参考文献
[1] CHANDRASEKHAR V, ANDREWS J, GATHER-ER A. Femtocell networks: a survey [EB/OL]. [2020-09-10]. https://arxiv.org/abs/0803.0952
[2] SHAFI M, MOLISCH A F, SMITH P J, et al. 5G: a tutorial overview of standards, trials, challenges, deployment, and practice [J]. IEEE journal on selected areas in communica-tions, 2017, 35(6): 1201-1221. DOI:10.1109/jsac.2017.2692307
[3] LIU J Y, SHENG M, LIU L, et al. Interfer-ence management in ultra-dense networks: challenges and approaches [J]. IEEE net-work, 2017, 31(6): 70-77. DOI:10.1109/mnet.2017.1700052
[4] WU J, ZHANG Z F, HONG Y, et al. Cloud radio access network (C-RAN): a primer [J]. IEEE network, 2015, 29(1): 35-41. DOI:10.1109/mnet.2015.7018201
[5] PAN C H, ELKASHLAN M, WANG J Z, et al. User-centric C-RAN architecture for ul-tra-dense 5G networks: challenges and methodologies [EB/OL].[2020-09-10]. https://arxiv.org/abs/1710.00790
[6] ZHENG J C, WU Y, ZHANG N, et al. Opti-mal power control in ultra-dense small cell networks: a game-theoretic approach [J]. IEEE transactions on wireless communica-tions, 2017, 16(7): 4139-4150. DOI:10.1109/twc.2016.2646346
[7] YANG C G, LI J D, NI Q, et al. Interfer-ence-aware energy efficiency maximization in 5G ultra-dense networks [J]. IEEE transac-tions on communications, 2017, 65(2): 728-739. DOI:10.1109/tcomm.2016.2638906
[8] GRONDMAN I, BUSONIU L, LOPES G A D, et al. A survey of actor-critic reinforcement learning: standard and natural policy gradi-ents [J]. IEEE transactions on systems, man, and cybernetics, part C (applications and re-views), 2012, 42(6): 1291-1307. DOI:10.1109/tsmcc.2012.2218595
[9] GHADIMI E, CALABRESE F D, PETERS G, et al. A reinforcement learning approach to power control and rate adaptation in cellular networks [EB/OL]. [2020-09-10]. https://arxiv.org/abs/1611.06497
[10] E-UTRA. Small cell enhancements for E-UTRA and E- UTRAN physical layer as-pects: TR37.840[S]. 3GPP, 2003
[11] 3GPP. Further advancements for E-UTRA physical layer aspects: TR 36.814 V9.2.0 [S]. 3GPP, 2017

作 者 简 介



秦爽，电子科技大学副教授；主要研究领域为移动及无线通信网络；先后主持和参与各类科研项目 20 余项；已发表论文 70 余篇。

董星辰，南京船舶雷达研究所助理工程师；研究方向为超密集网络智能覆盖增强技术。

冯钢，电子科技大学教授、博士生导师；主要研究领域为移动机无线通信网络；先后主持和参加各类科研项目 30 余项；发表学术论文 200 余篇。

上接第 38 页

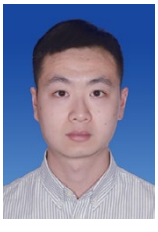
[5] JOUPPI N P, YOUNG C, PATIL N, et al. In-datacenter performance analysis of a tensor pro-cessing unit [C]//The 44th Annual International Symposium on Computer Architecture. Toron-to, Canada: ISCA, 2017
[6] GAMATIE A, DEVIC G, SASSATELLI G, et al. Towards energy-efficient heterogeneous multicore architectures for edge computing [J]. IEEE access, 2019, 7: 49474-49491. DOI: 10.1109/ACCESS.2019.2910932
[7] 肖汉，李彩林，李琦，等 . CPU+GPU 异构并行的矩阵转置算法研究 [J]. 东北师大学报 (自然科学版), 2019, 51(4): 70-77
[8] Gentsch P. AI business: framework and matu-rity model [M]. 2019
[9] WANG L, GUO S, HUANG W L, et al. Plac-es205-VGGNet models for scene recognition [J]. Computer science, 2015
[10] GIRSHICK R. Fast R-CNN [J]. Computer sci-ence, 2015. DOI: 10.1109/ICCV.2015.169
[11] 中国信息通信研究院，国家广播电视总局广播电视科学研究院，中国新闻出版传媒集团有限公司，等 . 云游戏产业发展白皮书 [R/OL]. (2019-12)[2020-08-07]. http://www.199it.com/ar-chives/988193.html
[12] 华为技术有限公司 . 5G 应用立场白皮书 [R/OL]. [2020-08-07]. http://www.lddoc.cn/p-12956741.html
[13] 唐洁，刘少山 . 面向无人驾驶的边缘高精地图服务 [J]. 中兴通讯技术，2019, 25(3): 58-67+81. DOI: 10.12142/ZTETJ.201903009

作 者 简 介

李建飞，中国联通研究院高级工程师；主要从事算力网络、AI 算法应用以及智能边缘计算的研究；发表论文多篇，获授权专利 10 余项。

曹畅，中国联通博士后、高级工程师，中国联通网络技术研究院未来网络研究部高级专家、智能云网技术研究室主任，第七届中国通信学会信息通信网络技术委员会委员，中国通信标准化协会网络 5.0 技术标准推进委员会架构组副组长，SDN/NFV 产业联盟 SDN 集成与互通测试组副组长，边缘计算网络基础设施联合工作组（ECNI）技术规范组组长；主要从事 IP 网络宽带通信、SDN/NFV、新一代网络编排技术的研究；发表论文 20 余篇，获授权专利 10 余项。

李奥，中国联通研究院助理工程师；主要从事人工智能在网络边缘应用的研究。



庞博文，中国联通研究院网络技术研究部助理工程师；主要从事数据分析、数据挖掘、网络人工智能等方面的研究。