DOI:10.19651/j.cnki.emt.1802009

数据中心网络 coflow 调度机制结构构建及仿真

李维虎[1] 张顶山[1] 崔慧明[1] 周 龙[1] 朱志挺[1] 谢 挺[2]
(1. 国网白银供电公司 白银 730900; 2.国家数字交换系统工程技术研究中心 郑州 450002)

摘 要: 重新构建得到了一种 coflow 调度算法-DeepCS,将 coflow 资源视图看成是需要进行后续处理的图像,根据之前学习策略来达到 coflow 的最佳调度效果。利用 DNN 提取特征参数时不必通过人为手动的方法进行设计,通过单独学习过程便可实现,给出深度增强学习系统。训练输入包含了各项网络与任务情景,并以动作概率分布作为输出,EPiSOdE 作为单位开展训练过程。仿真结果得到:当 coflow 到达速率变大后,将会导致所有算法需要更长的 coflow 完成时间,此时调度算法流时间与的工作压力都会增加,从而形成更长的 coflow 平均完成时间;在较低的 coflow 到达速率下,VARYS 和 DeepCS 具有相似的性能,都比 PFABRiC 的性能更好,并且 DeepCS 性能提升最快。

关键词: 数据中心;网络;语义相关流;调度机制;性能
中图分类号: TN914 **文献标识码**: A **国家标准学科分类代码**: 41320

Data center network coflow scheduling mechanism structure construction and simulation

Li Weihu[1] Zhang Dingshan[1] Cui Huiming[1] Zhou long[1] Zhu Zhiting[1] Xie Ting[2]
(1.State Grid Silver Power Supply Company, Baiyin 730900, China;
2.National Digital Exchange System Engineering Technology Research Center, Zhengzhou 450002, China)

Abstract: DeepCS, a kind of coflow scheduling algorithm, is obtained through reconstruction. The coflow resource view is regarded as the image to be processed later, and the optimal scheduling effect of coflow is achieved according to the previous learning strategy.The feature parameters extracted by DNN need not be designed by manual method, and can be realized by separate learning process. The training input includes various network and task situations, with the motion probability distribution as the output, and EPiSOdE as the unit to carry out the training process.The simulation results are as follows: when the coflow reaches a larger rate, all algorithms will need longer coflow completion time. At this point, the flow time and working pressure of the scheduling algorithm will increase, thus forming a longer average coflow completion time.Under the lower coflow arrival rate, VARYS and DeepCS have similar performance, both of which are better than PFABRiC, and DeepCS has the fastest performance improvement.

Keywords: data centres; network; semantic correlation flow; scheduling mechanism; performance

0 引 言

目前,数据中心网络(data centres network, DCN)在处理大部分计算任务时都会使用分布式数据处理系统。对这些系统进行流量分析可以发现:当系统接收到计算任务时会将其分解为众多的子任务,之后再将这些子任务分配给相应的主机,同时由不同主机生成的中间数据也会互相进行交换而成为后续子任务作为其它主机的输入数据[1-2]。邓有林[3]提出基于多维资源协调聚合的分组遗传资源挖掘算法,实现数据中心资源的均衡分配,该算法能够提升数据中心资源综合利用率,运行效率较高,可完成资源的均衡分配。

对于上述计算模式,当其中某一条数据流无法完成时,便会导致后一阶段的计算任务不能进行。对计算过程的各个环节时间进行关统可知,传输中间数据需要花费较长的时间,占到了所有计算任务总时间的近 30%,最大时间可以占到 50%[4-5]。曹晓峻等[6]提出数据网络与电力网络混合运行模型,建立以数据负荷为识别参数的数据中心能耗模型以及数据网络电力调节潜力分析模型。系统调控中心对数据网络下达电力、电量调度的具体要求,由数据网络进行数据负荷的转移响应。为了提高计算效率,必须采用合适的方法来优化上述数据传输过程,这也更有助于提升

收稿日期:2018-07

DCN 应用程序整体运算性能[7]。现阶段,人工智能领域已经在运算速度的提升方面取得了较大的突破,主要通过增强学习与深度学习的模式来促进学习效率的快速提高,并且以此方法作为数据运算处理核心的 AlphaGo 已经战胜了众多的人类围棋高手,也因此引起了全世界各领域的密切关注[8-10]。同时,深度增强学习方法也在通信网络领域开始获得广泛应用,对于改善网络资源管理效率发挥了重要作用[11-12]。胡智尧等[13]在终端电脑或者网络交换机上实现调度算法,从独立数据流调度方法和网络流组的调度方法进行分析,并进行展望。马腾等[14]提出一种新的语义相关流调度机制,仿真得到该调度机制均使得语义相关流的平均完成时间小于其他调度机制,尤其是网络负载较大时,性能提升约 50%。黄鸿等[15]着眼于先验知识未知的数据中心网络并行计算应用场景,提出了一种基于端口聚合流量的 coflow 调度机制 CSPAT,用于最小化平均 coflow 完成时间,得到该模型可以方便地部署在现有数据中心网络节点上,有效减小并行计算应用通信阶段的平均 coflow 完成时间。

在此基础上,本文重新构建得到了一种 coflow 调度算法——DeepCS。该算法可以将 coflow 资源视图看成是需要进行后续处理的图像,以此实现把带宽约束的 coflow 调度问题变成时间连续学习过程,根据之前学习策略来达到 coflow 的最佳调度效果。

1 coflow 调度机制

1.1 深度增强学习系统

在策略选择过程中,智能体通常会将系统策略 $\pi(S,A):S\times A\rightarrow[0,1]$ 定义成一个状态,在动作对与概率间形成的映射代表处于状态 S 下对动作 A 进行选择的概率,并符合条件 $\sum A\in A\pi(S,A)=1$。对实际问题进行处理时,会涉及多个动作对,因此增强学习必须具备良好的泛化能力,可以根据已知的有限学习经验与记忆内容对更大范围知识进行推断分析[16-17]。上述处理过程通常会选择策略函数来实现函数逼近的过程,可将其理解成是通过参数化函数 $\pi\theta(S,A)$ 来逼近状态动作对与概率间的映射关系,此处的 θ 代表策略参数。利用深度神经网络(deep neural network,DNN)提取特征参数时不必通过人为手动的方法进行设计,只需通过单独学习过程便可实现。本文把 DNN 应用于增强学习框架中,得到了深度增强学习系统,结果见图 1。

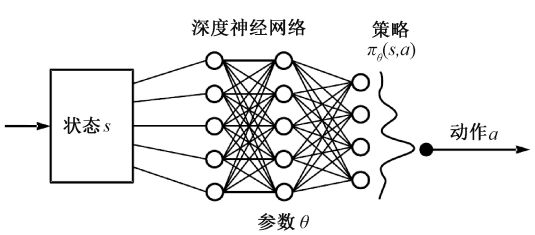

图 1 智能体的构成

策略梯度算法属于一种直接逼近方式的优化过程,是计算最优策略的最佳方法。开展增强学习的目的是为了最大程度地累积折扣奖赏,将其表示成估值函数的形式,可以将策略梯度表示为式(1):

$$\nabla_\theta\left(\sum_{t=0}^{\infty}\lambda_t\right)=\nabla_\theta\log\pi_\theta(s,a) \tag{1}$$

式中:$\pi_\theta(s,a)$ 是利用处于 S 状态下的策略 π_θ 时通过动作 a 得到的累积折扣奖赏。此算法的关键内容是根据现有策略对运行轨迹进行分析得到的梯度估计结果。采用蒙特卡洛计算方法时,可以通过智能体来完成轨迹的多点采样过程,并计算出累积折扣奖赏 v_t,再将其作为 $\pi_\theta(s,a)$ 无偏估计,之后根据式(2)更新策略参数 θ:

$$\theta\leftarrow\theta+\alpha\sum_t\nabla_\theta\log\pi_\theta(S_t,a_t)v_t \tag{2}$$

式中:t 代表各个迭代过程的步进方向,实际步进值取决于参数 v_t。利用 REiNFORCE 算法获得的梯度估计属于无偏估计,不过采用此方法计算获得的结果具有较大方差,不利于算法收敛,极大降低了算法效率。本文主要利用回报基线法进行处理,先设置一个极小变量,该变量可以通过截取 VT 返回值来获得,使梯度估计方差显著减小。

1.2 训练过程

训练输入包含了各项网络与任务情景,并以动作概率分布作为输出,选择幕(EPiSOdE)作为单位开展训练过程。对各幕的固定任务到达网络之后再按照策略函数实施调度,当执行完各项任务后,则幕结束。

为确保训练结果达到一般性条件,接收到集中训练任务时需对各个随机过程进行遍历操作。同时,在实际训练过程中,各迭代过程都由 N 个幕组成,所有幕都利用随机任务到达时序作为训练集,并测试在现有策略函数下通过动作空间得到的概率分布,同时选择探测到的提升策略函数作为不同训练集的训练结果。各迭代过程都是以任务到达时序具有各自随机特性的训练集。同时对所有迭代期间的各幕系统状态进行数据记录,根据上述记录数据推导出 T 时刻对应的幕累积奖赏 VT。结合以上数据,并根据之前构建的式(2)便可利用迭代的方式来完成训练的目的。

采用 Reinforce 算法估计梯度时,存在方差较大、算法收敛慢的弊端,为克服这种弊端,一般采用回报基线法,截取返回值。回报基线法形式多种多样,本文采取的办法是对相同训练集下各幕相同时刻的返回值求取平均数。

2 性能评估

2.1 系统结构

在 DeepCS 系统中总共包括了 2 个组成部分:1)集中调度器,2)具有分布式特点的端系统执行组件[18-19]。该系统的具体架构如图 2 所示,考虑到实际拥塞只会出现于边缘链路中,DeepCS 不会在网络交换机中构建功能模块。在所有的物理服务器中都存在端系统执行组件作为守护程序,其作用是发送各类 coflow 信息以及链路资源数据

到调度器中。当准备好发送端 coflow 数据后，再通过守护程序将 coflow 信息报告给 coflow 收集模块；当接收端各项准备工作就绪后，由中心调度器按照实际网络与任务情景通过深度增强学习法给出最优调度策略，计算出 coflow 传输过程需消耗的时间及数据发送速度，再把结果发送至端系统的执行组件。通过端系统的执行组件确定中心调度器起始时间与数据发送速度后再开启 coflow 进行数据传输。

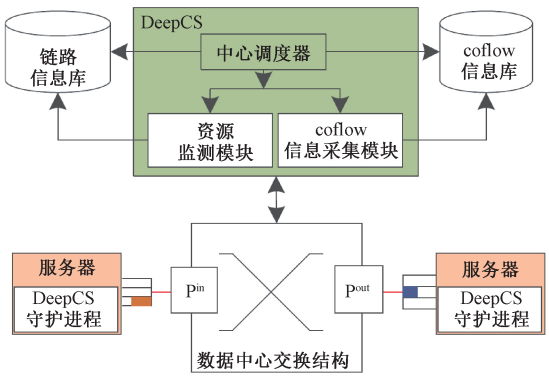

图 2　DeepCS 系统架构

2.2　仿真环境搭建

利用 NS3 开展包级仿真[20]，并对 DeepCS 性能进行测试。测试过程的所有流量数据都是根据产业界数据中心网络提供的 coflow 流量得到。分别从宽度与容量两个不同维度对 coflow 进行描述。测试期间的 coflow 变化范围是 10～100，宽度变化范围介于 1～50，并且宽度在 16 以下的 coflow 比例是 60%。在交换网络中，coflow 数量分布规律表现为 $\lambda \in [0.2, 0.8]$ 泊松分布的特点，可以模拟出流量动态变化的趋势。利用非阻塞交换网络在 32 台服务器之间构建数据传输通道，链路带宽等于 1 Gbps。流量的发送与接收端都是通过轮询的形式在边缘链路上形成均匀分布状态，并且当 coflow 改变后，数量也会随之发生变化，不过必须确保 80%左右属于 coflow。

对系统进行仿真测试的对比算法包括：

1）TCP 基线算法：在网络中传输的各个数据流都是通过 TCP 协议进行带宽的公平竞争，从而确保 MAX 与 MIN 满足公平性条件。

2）PFABRiC 算法：是对单个流进行调度的最佳算法，对于无阻塞的交换网络，需选择 PFABRiC 理想变体实施对比分析。

3）VARYS 算法：属于一种 coflow 级以优先级为基础的调度方法，其核心内容为最小瓶颈优先策略，属于现阶段具有最佳性能的 coflow 级调度算法。

利用网络仿真器 NS3 比较了 DeepCS、TCP、PFAB、RiC、VARYS 各种调度算法处理流所需的平均时间和网络负载及复用因子间的关系。对比分析方法 1 与方法 2 的性能优势并将其定义成（CCT2－CCT1）/CCT2，采用这两种方法完成 coflow 的平均时间依次为 CCT1 与 CCT2，并以相对基线算法的性能提高值作为本文性能评价指标。

2.3　算法性能对比

各个网络负载的 coflow 平均完成时间都把复用因子设定成 0.2，并将 coflow 到达速率 λ 分别确定为 0.2、0.4、0.6 与 0.8，由此模拟得到各个网络负载。利用网络仿真器 NS3 比较了不同调度算法所需的 coflow 平均完成时间，结果如图 3 所示。

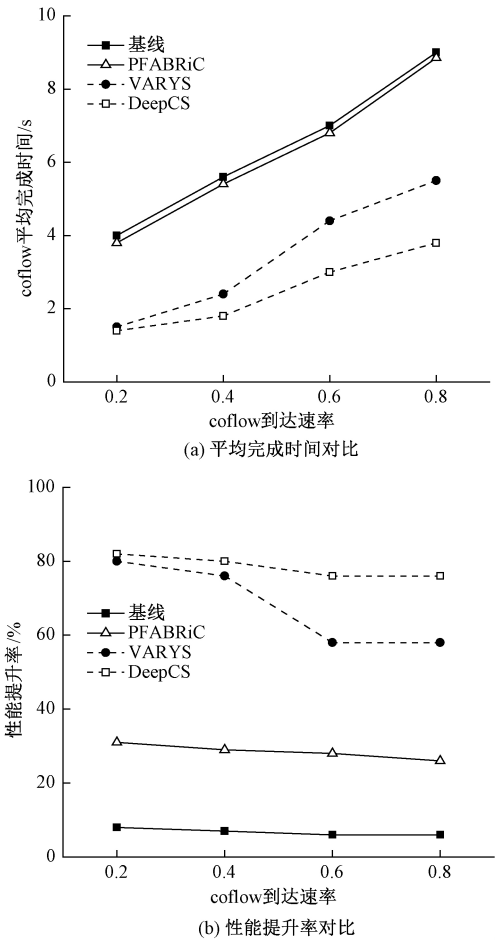

图 3　不同网络负载的平均流完成时间及性能提升对比

根据图 3 可知：1）当 coflow 到达速率变大后，将会导致所有算法需要更长的 coflow 完成时间，此时调度算法流时间与工作压力都会增加，从而形成更长的 coflow 平均完成时间；2）DeepCS 具有比其它调度算法更佳的 coflow 平均完成时间，这是由于 DeepCS 在应用之前对各类流量模式进行了全面学习，从而得到最优调度策略；3）在较低的 coflow 到达速率下，VARYS 和 DeepCS 具有相似的性能，都比 PFABRiC 的性能更好，并且 DeepCS 性能提升最快。

3 结 论

利用 DNN 提取特征参数时不必通过人为手动的方法进行设计,通过单独学习过程便可实现,给出深度增强学习系统。训练输入包含了各项网络与任务情景,并以动作概率分布作为输出,EPiSOdE 作为单位开展训练过程。对各幕的固定任务到达网络之后再按照策略函数实施调度,当执行完各项任务后,则幕结束。

当 coflow 到达速率变大后,将会导致所有算法需要更长的 coflow 完成时间,此时调度算法流时间与的工作压力都会增加,从而形成更长的 coflow 平均完成时间;在较低的 coflow 到达速率下,VARYS 和 DeepCS 具有相似的性能,都比 PFABRiC 的性能更好,并且 DeepCS 性能提升最快。

参考文献

[1] 樊自甫,张丹,李书.基于软件定义网络的数据中心网络负载均衡算法研究[J].计算机工程与科学,2018,40(6):1017-1022.

[2] 钱伟强.基于海量存储云调度机制的云网络数据存储算法[J].国外电子测量技术,2017,36(3):27-30.

[3] 邓有林.大型 Web 网络数据中心资源高效挖掘技术研究[J].现代电子技术,2018,41(3):120-123.

[4] 王跃飞,曹三峰,毕翔,等.一种基于时隙动态分配的 FlexRay 系统通信机制[J].电子测量与仪器学报,2015,29(2):179-186.

[5] 高赐威,曹晓峻,闫华光,等.数据中心电能管理及参与需求侧资源调度的展望[J].电力系统自动化,2017,41(23):1-7.

[6] 曹晓峻,高赐威,李德智,等.数据网络与电力网络混合运行建模及其参与系统经济运行[J].中国电机工程学报,2018,38(5):1448-1456.

[7] 樊自甫,李书,张丹.基于流量调度的 SDN 数据中心网络拥塞控制算法[J].计算机科学,2017,44(Z1): 266-269,273.

[8] 马铭冀,张晓蕾,杨继家.云时代下数据中心网络技术研究[J].科技创新与应用,2017(15):99.

[9] 黄忠建,代红兵,王蕾.嵌入式 Forth 操作系统实时调度算法研究[J].计算机应用研究,2019(10):1-2.

[10] 许文庆, 余庚. SDN 架构下数据流量调度算法的设计[J].光通信研究,2018(3):5-8,20.

[11] 陈睿,庞海萍,郝丽,等.Powerlink 协议异步调度机制的建模与分析[J/OL].计算机工程与应用:1-10. http://kns. cnki. net/kcms/detail/11.2127. tp. 20181101.0950.013.html.

[12] 邹云峰,陈宇,邱文玮,等.边缘计算环境下服务质量感知的资源调度机制[J].电子技术与软件工程,2018(18):178-179.

[13] 胡智尧,李东升,李紫阳.数据中心网络流调度技术前沿进展[J].计算机研究与发展,2018,55(9):1920-1930.

[14] 马腾,胡宇翔,张校辉.基于深度增强学习的数据中心网络 coflow 调度机制[J].电子学报,2018,46(7):1617-1624.

[15] 黄鸿,莫李思,孙罡.一种基于端口聚合流量的 Coflow 调度机制[J].通信技术,2018,51(7):1594-1601.

[16] 蔡震震,唐鹏,胡建斌,等.深度卷积神经网络实现硬性渗出的自动检测[J].计算机科学,2018,45(Z2):203-207.

[17] 郭智,宋萍,张义,等.基于深度卷积神经网络的遥感图像飞机目标检测方法[J].电子与信息学报,2018,40(11):2684-2690.

[18] 郝春亮,沈捷,张珩,等.大数据背景下集群调度结构与研究进展[J].计算机研究与发展,2018,55(1):53-70.

[19] 郑振,乔庐峰,陈庆华,等.星载 IP 交换机中变长调度 Clos 交换结构的设计[J].通信技术,2016,49(3):361-367.

[20] 王珊珊,季海波,司鹏,等.基于 NS3 的 LTE 网络圆形边界下移动模型的仿真[J].小型微型计算机系统,2015,36(11):2531-2535.

作者简介

李维虎,本科,高级工程师,主要研究方向为电力生产管理。

E-mail:caoqunfeng811646@126.com