计算机科学
COMPUTER SCIENCE

基于BiLSTM神经网络的多服务器门限服务系统性能分析

杨志军, 黄文洁, 丁洪伟

引用本文

杨志军, 黄文洁, 丁洪伟. 基于BiLSTM神经网络的多服务器门限服务系统性能分析[J]. 计算机科学, 2023, 50(10): 266-274.

YANG Zhijun, HUANG Wenjie, DING Hongwei. Performance Analysis of Multi-server Gated Service System Based on BiLSTM Neural Networks [J]. Computer Science, 2023, 50(10): 266-274.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000221

基于 BiLSTM 神经网络的多服务器门限服务系统性能分析

杨志军[1,2,3]　黄文洁[1]　丁洪伟[1]
1 云南大学信息学院　昆明 650500
2 云南省教育厅教学仪器装备中心　昆明 650223
3 云南师范大学教育部民族教育信息化重点实验室　昆明 650500

摘　要　为了满足运行速度快、时延低、性能好、公平性好等特点，提出了多服务器门限服务系统，并利用BiLSTM(Bi-directional Long Short-Term Memory)神经网络对其进行预测分析，使用多服务器接入方式来降低网络时延，改善系统性能。多个服务器调度时，可以采用同步和异步两种方式。首先，研究多服务器门限服务的系统模型。其次，在单服务器的基础上，利用嵌入马尔可夫链和概率母函数的分析方法对多服务器门限服务的平均排队队长、平均循环周期和平均时延进行求解；同时，利用Matlab进行仿真实验，分别将单服务器系统与多服务器系统的理论值与仿真值进行系统分析，对比多服务器同步和异步两种方式。最后，构建BiLSTM神经网络来预测多服务器系统的性能。实验结果表明，该多服务器系统异步方式优于同步和单服务器系统，多服务器异步系统的性能更好，时延更低，效率更高。综合对比多服务器的3种基本服务系统，在保证公平性的情况下，门限服务系统更加稳定。并且使用BiLSTM神经网络预测算法能够准确预测系统的性能，提高计算效率，对轮询系统的性能评价具有指导意义。

关键词：多服务器；同步方式；异步方式；平均排队队长；平均时延；公平性；BiLSTM神经网络
中图法分类号　TN911

Performance Analysis of Multi-server Gated Service System Based on BiLSTM Neural Networks

YANG Zhijun[1,2,3], HUANG Wenjie[1] and DING Hongwei[1]
1 School of Information Science and Technology, Yunnan University, Kunming 650500, China
2 Educational Instruments and Facilities Service Center, Educational Department of Yunnan Province, Kunming 650223, China
3 Key Laboratory of Education Informatization for Nationalities of Ministry of Education, Yunnan Normal University, Kunming 650500, China

Abstract　In order to meet the requirements of fast operation, low delay, good performance and fairness, a multi-server gated service system is proposed and its predictive analysis is carried out using BiLSTM (bi-directional long short-term memory) neural networks. Multi-server access is used to reduce the network delay and improve system performance. Both synchronous and asynchronous approaches can be used when multiple servers are scheduled. Firstly, the system model of multi-server gated service is investigated. Secondly, the average queue length, average cycle period and average delay of multi-server gated service are solved on the basis of single server using the analytical methods of embedded Markov chain and probabilistic generating function. Meanwhile, simulation experiments are conducted using Matlab to compare the theoretical and simulated values of single server system and multi-server system respectively system analysis, comparing both multi-server synchronous and asynchronous approaches. Finally, a BiLSTM neural network is constructed to predict the performance of the multi-server system. Experiments show that the asynchronous approach of this multi-server system is superior to the synchronous and the single-server system, and the multi-server asynchronous system has better performance, lower delay and higher efficiency. Comparing the three basic multi-server service systems, the gated service system is more stable while ensuring fairness. And the use of BiLSTM neural network prediction algorithm can accurately predict the performance of the system and improve the computational efficiency, which is a guideline for the performance evaluation of the polling system.

Keywords　Multi-server, Synchronous mode, Asynchronous mode, Average queue length, Average delay, Fairness, BiLSTM neural network

到稿日期：2022-10-25　返修日期：2023-02-18
基金项目：国家自然科学基金(61461053,61461054)；“云南省兴滇英才支持计划”产业创新人才专项(YNWR-CYJS-2020-017)
This work was supported by the National Natural Science Foundation of China(61461053,61461054) and Industry Innovation Talents Project of Yunnan Xingdian Talents Support Plan(YNWR-CYJS-2020-017).
通信作者：杨志军(353738698@qq.com)

5G 时代，网络规模逐渐扩大，传统单服务器轮询系统使用不同的服务策略为用户提供服务。由于单服务器的 CPU、内存等有很大的局限性，当网络业务量较大时，系统性能降低，运行速度变慢，不能满足数据实时传输的要求。随着互联网的发展，对密集站点数据的采集和传输，要求轮询系统的服务效率更高，时延更低。此外，由于轮询系统的三大特征，即公平性、高效性、高服务质量性，其被广泛应用到无线局域网中。

近年来，为了提高系统的性能和效率，降低时延，研究者们做了无数的尝试。他们发现，解决这类问题归根结底有两种方法：第一种方式是提升系统服务器的性能，配置更高级的硬件设备，如更好的 CPU 和更大容量的内存等；第二种方法是增加服务器的数量，利用多服务器处理数据的方式来提升系统的性能和效率[1-2]。两种方法都具有可行性，但是由于配置硬件设备成本较高，不易于推广使用，因此在实际的处理过程中，一般采用增加服务器数量的方法来解决此类问题。文献[3]论述了轮询系统在计算机网络通信等领域的应用。文献[4]提出了一种双队列混合服务策略轮询系统，用于实现不同的接入方式，采用不同的轮询策略，间接地提高了系统的效率。文献[5]提出了多队列双服务器同步控制策略轮询系统，利用双服务器同步调度方式来提高处理效率，降低时延。文献[6]利用双服务器的可伸缩结构来分析双服务器异步控制策略轮询系统的性能，提高了系统的兼容性和效率。文献[7-8]通过局域网的连接和接入服务器的管理，将多台服务器组成单站点的服务器集群，将用户的访问请求分配给集群中的每台服务器，有效提高了系统的可扩展性和容错能力。

在上述文献中，对多服务器轮询系统的研究还较少，大多数都是基于改进服务策略来改善系统的性能。本文提出的多服务器门限服务系统能够在很大程度上弥补单服务器带来的缺陷，提高服务效率，并且在降低时延、提高效率的基础上保证公平性。此外，密集站点的接入和大量数据的获取使网络维护部署越来越困难，故找到一种准确可靠的神经网络预测轮询系统性能的方法，对网络的维护、评价具有指导意义。因此在此基础上，本文利用 BiLSTM 神经网络预测多服务器门限服务系统。

目前，深度学习已被广泛应用到预测分析中，如卷积神经网络(Convolutional Neural Networks，CNN)、循环神经网络(Recurrent Neural Network，RNN)、BP(Back Propagation)神经网络、BiLSTM(Bi-directional Long Short-Term Memory)神经网络等，利用前期数据预测后期数据及趋势。文献[9]利用 BiLSTM 与注意力机制，准确预测了设备的剩余使用寿命。文献[10]利用特征筛选的卷积神经网络与双向长短期记忆网络组合的模型对多维的短期电力负荷进行预测，有效提高了预测精度。

本文将轮询系统的相关知识应用到平均队长、时延、周期这三大指标的分析过程中，通过仿真实验进行合理验证，并且利用 BiLSTM 神经网络对多服务器轮询系统进行预测分析。最后，实验结果表明，理论值与仿真值、理论值与预测值都基本一致，该方案可行，表明增加服务器数量能够有效提高系统的效率，降低时延。

1 背景知识

1.1 轮询系统

轮询系统的模型可表示为由 N 个站点和 1 个服务器组成，如图 1 所示。服务器依据一定规则从一个方向为每个站点提供服务，在实际过程中一般按照站点编号的升序方向进行询问并服务。待最后一个站点发送服务请求并且服务完成，再返回第一个站点进行服务。在实际应用中，根据单服务器或多服务器模型选取一个或者多个逻辑上的中心按照一定的周期顺序对各个站点进行询问，进而为发送服务请求的站点提供服务[11]。

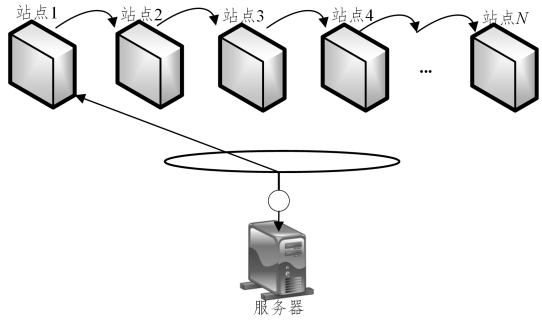

图 1 轮询系统模型

Fig. 1 Polling system model

轮询系统按照服务策略通常分为门限、完全、限定。门限服务系统指服务器仅对服务站点时刻之前到达的信息分组提供服务。完全服务系统指服务器为某一个站点提供服务直到该站点为空，转入下一个站点进行服务。限定服务系统指询问到任意一个站点时，每次最多为 K 个信息分组提供服务。3 种不同的服务策略各有特点，从平均时延来说，完全服务优于门限服务，门限服务优于限定服务；从公平性来说，限定服务优于门限服务，门限服务优于完全服务；综合平均时延和公平性来说，门限服务最适合。轮询系统根据服务器数量又可分为单服务器轮询系统和多服务器轮询系统，其中针对业务量较大的情况，单服务器时延较高，效率较低。因此，本文提出了一种多服务器接入方式。

1.2 无线局域网

无线局域网(Wireless Local Area Network，WLAN)具有大容量性、抗干扰性、灵活性、易扩展性等特点，不能简单地使用有线局域网的协议。由于无线局域网的传输条件特殊，造成信号强度的动态范围极大，让发送站很难利用冲突检测方法来确定是否发生冲突[12]。在 IEEE 802.11 标准中采用 CSMA/CA 基本多址技术，CA 指避免冲突，该协议用避免冲突检测方法来代替在 802.3 协议中使用的冲突检测。载波检测 CSMA 通过两个机制共同完成，分别是物理机制和虚拟机制。物理载波机制由物理层提供，虚拟机制由 MAC 提供[13]。物理载波机制使用的是信道空闲评估 CCA 算法，CCA 是基于一个给定阈值上的能量检测和基于相关质量码字锁定的检测，它允许站点共用一个媒体，每次只允许一个节点发送信息，每个用户在发送数据之前都要侦听信道，直到媒体空闲才可以发送，当两个站点都在媒体空闲时同时发送数据，容易发生碰撞和竞争。因此，为了避免冲突，IEEE 802.11 标准设计

了一个独特的 MAC 子层，它包括两个功能：1)分布式协调功能 DCF；2)点协调功能 PCF[14-15]。WLAN 中轮询机制通常采用门限服务的接入控制策略，即获得传送权的站点只发送完服务时限(Service Interval，SI)内到达的信息分组数，传送期间到达的信息分组将在下一次获得传送权时发送，如图 2 所示。

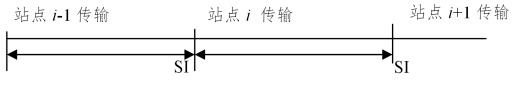

图 2 轮询系统间隔发送示意图

Fig. 2 Schematic diagram of polling system interval sending

1.3 多服务器轮询系统

多服务器轮询系统由 N 个站点和 S 个服务器组成，如图 3 所示。多服务器的调度方式可分为两种：同步方式和异步方式。同步方式指当某一个站点发送服务请求时，由 S 个服务器同时为该站点服务，直到所有服务器服务完成才进行下一个站点的询问，同步方式服务状态如图 4 所示。异步方式指当某一个站点发送服务请求时，由一个服务器为其提供服务，待该服务器结束服务后，可进入下一个站点进行询问。也就是说，当一个服务器为站点 i 提供服务时，另一个服务器直接查询 $i+1$ 站点[16]。异步方式服务状态如图 5 所示。

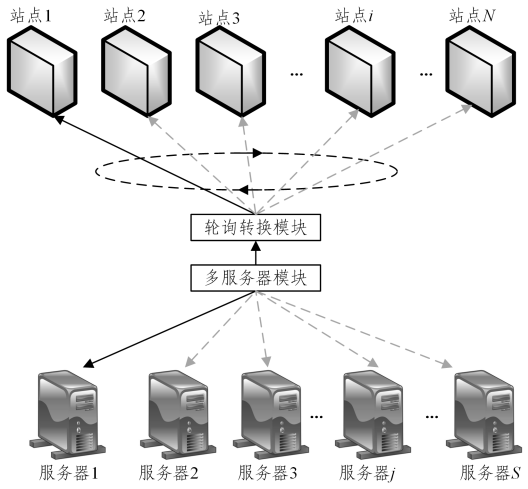

图 3 多服务器轮询系统模型

Fig. 3 Multi-server polling system model

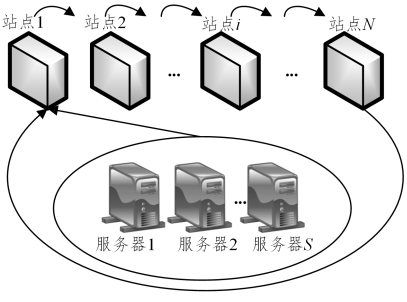

图 4 同步方式服务状态图

Fig. 4 Service status diagram in synchronous mode

通过分析同步和异步两种方式可以得出，在单服务器轮询系统中，两种服务策略的效果是一样的。各个站点发送服务请求时，服务器按照门限服务的规则进行服务。在多服务器轮询系统中，采用异步方式进行服务，即当有多个站点发送服务请求时，系统可将多个站点的服务请求分别发送到不同的服务器上，再将结果反馈到用户端，大大提高了效率。当站点发送服务请求后，服务请求先被服务器模块接收到，多服务器模块将请求站点的 MAC 地址转换为目标服务器的 MAC 地址，然后将服务请求发送出去，目标服务器即可为该站点提供服务[17]。

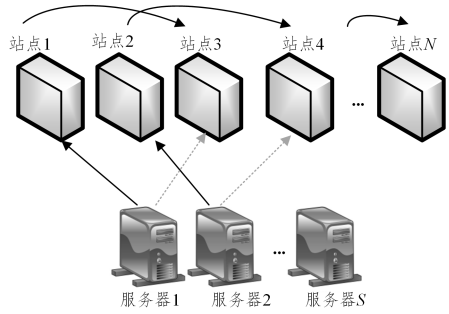

图 5 异步方式服务状态图

Fig. 5 Service status diagram in asynchronous mode

2 系统分析

探究系统的性能一般使用系统的一阶特性和二阶特性。一阶特性指平均队长和周期，二阶特性指平均时延。在相同条件下，平均队长、周期和时延越小，表明系统的性能越好，运行速度越快。因此，我们定义随机变量和概率母函数来分析多服务器的特征参数[18]。

2.1 系统工作条件

根据轮询系统的工作机制和特点、通信过程中数据的到达过程和传输特性，对系统工作条件作如下定义[19-20]：

1)进入各个站点内等待发送的信息分组的到达过程是相互独立的，是同分布的泊松分布，其分布的概率母函数为 $A(z)$，均值为 $\lambda=A'(1)$，方差为 $\sigma_\lambda^2=A''(1)+\lambda-\lambda^2$。

2)服务器按门限服务规则对站点内的信息分组后提供服务的时间是相互独立且同分布的，其分布的概率母函数为 $B(z)$，均值为 $\beta=B'(1)$，方差为 $\sigma_\beta^2=B''(1)+\beta-\beta^2$。

3)任意两个相邻站点之间的查询转换时间服从相互独立、同分布的随机变量，其概率母函数为 $R(z)$，均值为 $\gamma=R'(1)$，方差为 $\sigma_\gamma^2=R''(1)+\gamma-\gamma^2$。

4)轮询系统中每一个站点都有足够大的存储器容量，能够保证信息分组不丢失。

5)服务器将会按照先发送先服务的原则(FCFS)为站点提供服务。

6)多服务器系统中，站点数为 $N(N\geqslant 1)$，服务器数为 $S(S\geqslant 1)$。

7)轮询系统的查询顺序为服务器先按照站点编号的升序进行询问，直到询问完最后一个站点 N，又返回站点 1 进行询问。

2.2 随机变量

假设 i 号站点在 t_n 时刻，由 S 个服务器按照门限服务规则的同步控制方式为其服务，当 i 号站点发送完其存储内的信息分组后，即 S 个服务器对站点 i 服务结束后继续查询

下一个站点 $i+1$，$i+1$ 号站点在 t_{n+1} 时刻被服务。设定随机变量 $\xi_i(n)$ 为站点 i 在 t_n 时刻的存储的信息分组数，整个排队系统可以表示为 $[\xi_1(n),\xi_2(n),\xi_3(n),\cdots,\xi_i(n),\cdots,\xi_N(n)]$，概率分布函数为 $\pi_i(x_1,x_2,x_3,\cdots,x_i,\cdots,x_N)$。随机变量 $\xi_i(n+1)$ 为在 t_{n+1} 时刻第 $i+1$ 号站点存储器内的信息分组数可表示为 $[\xi_1(n+1),\xi_2(n+1),\xi_3(n+1),\cdots,\xi_i(n+1),\cdots,\xi_N(n+1)]$，概率分布函数为 $\pi_i(y_1,y_2,y_3,\cdots,y_i,\cdots,y_N)$。

随机变量的定义如表 1 所列。

表 1　随机变量

Table 1　Random variables

名称	变量定义
$u_i(n)$	服务器从 i 号站点到 $i+1$ 号站点的查询时间
$v_i(n)$	服务器对 i 号站点进行门限服务的时间
$\mu_j(u_i)$	在 $u_i(n)$ 时间内进入 j 号站点内的信息分组数
$\eta_j(v_i)$	在 $v_i(n)$ 时间内进入 j 号站点内的信息分组数

2.3　概率母函数

为了构建系统的数学模型，引入了概率母函数的分析方法，系统在 $\sum_{i=1}^{N}\lambda\beta=N\lambda\beta=N\rho<S$ 的条件下达到稳定。其中，$\rho=\lambda\beta$ 表示负载，S 表示服务器数量。则：

$$\lim_{n\to\infty}P[\xi_i(n)=x_i;i=1,2,\cdots,N]=\pi_i(x_1,x_2,x_3,\cdots,x_i,\cdots,x_N)$$

概率母函数定义为：

$$G_i(z_1,z_2,z_3,\cdots,z_i,\cdots,z_N)=\sum_{x_1=0}^{N}\sum_{x_2=0}^{N}\sum_{x_3=0}^{N}\cdots\sum_{x_i=0}^{N}\cdots\sum_{x_N=0}^{N}\pi_i(x_1,x_2,x_3,\cdots,x_i,\cdots,x_N)\cdot z_1^{x_1}z_2^{x_2}z_3^{x_3}\cdots z_i^{x_i}\cdots z_N^{x_N},i=1,2,3,\cdots,N \tag{1}$$

当服务器在 t_{n+1} 时刻开始为 $i+1$ 号终端站提供服务时，有：

$$\begin{cases}\xi_j(n+1)=\xi_j(n)+\mu_j(u_i)+\eta_j(v_i)\\ \xi_i(n+1)=\mu_i(u_i)+\eta_i(v_i)\end{cases},j\neq i \tag{2}$$

由此可得状态变量在 t_{n+1} 时刻的概率母函数为：

$$\begin{aligned}&G_{i+1}(z_1,z_2,\cdots,z_i,\cdots,z_N)\\&=\lim_{n\to\infty}E[\prod_{j=1}^{N}z_j^{\xi_j(n+1)}]\\&=\lim_{n\to\infty}E[\prod_{\substack{j=1\\j\neq i}}^{N}z_j^{\xi_j(n)+\mu_j(u_i)+\eta_j(v_i)}\cdot z_i^{\mu_i(u_i)+\eta_i(v_i)}]\\&=\lim_{n\to\infty}E[\prod_{\substack{j=1\\j\neq i}}^{N}z_j^{\xi_j(n)+\eta_j(v_i)}\cdot z_i^{\eta_i(v_i)}]\cdot R_i[\prod_{j=1}^{N}A_j(z_j)]\\&=G_i[z_1,z_2,\cdots,z_{i-1},B_i(\prod_{j=1}^{N}A_j(z_j)),z_{i+1}\cdots,z_N]\cdot R_i[\prod_{j=1}^{N}A_j(z_j)]\\&=R_i[\prod_{j=1}^{N}A_j(z_j)]\cdot G_i[z_1,z_2,\cdots,z_{i-1},B_i(\prod_{j=1}^{N}A_j(z_j)),z_{i+1},\cdots,z_N],i=1,2,\cdots,N\end{aligned} \tag{3}$$

2.4　平均排队队长

定义在 t_n 时刻，服务器按照门限服务规则为第 i 号站点的信息分组提供服务时，第 j 号站点存储器中的平均存储的信息分组数为 $g_i(j)$，则有：

$$g_i(j)=\lim_{z_1,z_2,\cdots,z_i,\cdots,z_N\to1}\frac{\partial G_i(z_1,z_2,\cdots,z_i,\cdots,z_N)}{\partial z_j},\quad i,j=1,2,\cdots,N \tag{4}$$

通过式(3)和式(4)可计算得到多服务器门限服务系统的平均排队队长，具体为：

$$g_i(i)=\frac{NS\lambda\gamma}{S-N\lambda\beta} \tag{5}$$

2.5　平均循环周期

系统的平均循环周期指服务器对系统中的站点按照门限服务规则完成一次遍历所用的时间。由服务时间和轮询转换时间这两部分组成。

本文提出了一种新的计算方法，利用到达率和平均排队队长求得平均循环周期 $E(\theta)$。

$$\frac{g_i(i)}{E(\theta)}=\lambda\Rightarrow E(\theta)=\frac{g_i(i)}{\lambda}=\frac{NS\gamma}{S-N\lambda\beta} \tag{6}$$

2.6　平均时延

平均时延指信息分组从进入站点存储器到开始被发送出去的时间，由 3 部分组成，分别为 $E(w_{i,1})$，$E(w_{i,2})$，$E(w_{i,3})$。

二阶特性定义为：

$$g_i(j,k)=\lim_{z_1,z_2\cdots,z_j\cdots,z_k,\cdots,z_N\to1}\frac{\partial^2G_i(z_1,z_2,\cdots,z_j,\cdots,z_k,\cdots,z_N)}{\partial z_j\partial z_k},\quad i=1,2,\cdots,N;j=1,2,\cdots,N;k=1,2,\cdots,N \tag{7}$$

由式(7)可得：

$$\begin{aligned}g_{i+1}(j,k)=&\lambda^2R''(1)+\gamma\lambda^2+\gamma\lambda[g_i(k)+g_i(j)]+[\lambda\rho+2\gamma\lambda\rho+\lambda^2B''(1)]g_i(i)+g_i(j,k)+\rho[g_i(j,i)+g_i(i,k)]+\rho^2g_i(i,i),i\neq j\neq k\end{aligned} \tag{8}$$

$$\begin{aligned}g_{i+1}(j,j)=&\lambda_j^2R_i''(1)+\gamma_iA_j''(1)+2\lambda_j\gamma_ig_i(j)+[2\lambda_j^2\hat{\beta}_i\gamma_i+\hat{\beta}_iA_j''(1)+\lambda_j^2B_i''(1)]g_i(i)+g_i(j,j)+2\lambda_j\hat{\beta}_ig_i(j,i)+\lambda_j^2\hat{\beta}_i^2g_i(i,i),i\neq j\end{aligned} \tag{9}$$

$$\begin{aligned}g_{i+1}(j,i)=&\lambda_j\lambda_iR_i''(1)+\lambda_j\lambda_i\gamma_i+\lambda_i\gamma_ig_i(j)+[2\lambda_j\lambda_i\hat{\beta}_i\gamma_i+\lambda_j\lambda_i\hat{\beta}_i+\lambda_j\lambda_iB_i''(1)]g_i(i)+\lambda_i\hat{\beta}_ig_i(j,i)+\lambda_j\lambda_i\hat{\beta}_i^2g_i(i,i),i\neq j\end{aligned} \tag{10}$$

$$g_{i+1}(i,i)=\lambda_i^2R_i''(1)+\gamma_iA_i''(1)+\lambda_i^2\hat{\beta}_i^2g_i(i,i)+[2\lambda_i^2\hat{\beta}_i\gamma_i+\hat{\beta}_iA_i''(1)+\lambda_i^2B_i''(1)]g_i(i) \tag{11}$$

由上述表达式计算得：

$$\begin{aligned}g_i(i,i)=&\frac{N}{(1+\rho)(1-N\rho)}\{\lambda^2R''(1)+\frac{1}{1-N\rho}[(N+2N\rho-1)\lambda^2\gamma^2+(N-1)\lambda^2\rho\gamma+(1+\rho-N\rho)\gamma A''(1)+N\lambda^3\gamma B''(1)]\},i=1,2,\cdots,N\end{aligned} \tag{12}$$

由式(7)－式(12)可得平均时延为：

$$\begin{aligned}E(w_G)=&E(w_{i,1})+E(w_{i,2})+E(w_{i,3})\\=&\frac{(1+\rho)g_i(i,i)}{2\lambda g_i(i)}\\=&\frac{1}{2}\left\{\frac{R''(1)}{\gamma}+\frac{1}{1-N\rho}[(N-1)\gamma+(N-1)\rho+2N\gamma\rho+N\lambda B''(1)+\frac{(1+\rho-N\rho)A''(1)}{\lambda^2}]\right\}\end{aligned} \tag{13}$$

其中，$\hat{\beta}=\frac{\beta}{S}$，$\rho=\lambda\hat{\beta}=\lambda\frac{\beta}{S}$。

3 仿真实验

根据上述理论分析，利用 Matlab 搭建多服务器模型，采用同步和异步两种方式对该模型的门限服务策略进行实验仿真，从实验角度验证理论分析的合理性，评估系统的性能。通过设定参数得到平均队长、周期、时延的理论值和仿真值，对两者进行比较。在实验中，任意单位时隙内所有站点的信息分组都是按泊松分布规律到达的，并且都是相互独立、同分布的。假设在仿真实验中全部都是理想状态，信息传输是没问题的，且系统在 $\sum_{i=1}^{N}\lambda\beta=N\lambda\beta=N\rho<S$ 的情况下达到稳定状态。设定信息分组进入站点的到达率为 0.005～0.05，以步长为 0.005 依次增加。每一个信息分组到达站点接受服务器服务的时间为 2 时隙，服务器对站点的询问时间为 1 时隙。为了提高模型的准确性，仿真实验的循环次数设置为 100000。

随着到达率的变化，探究门限服务策略的队长、周期和时延；利用多服务器轮询系统的门限、完全、限定这 3 种服务策略，探究多服务器门限服务的优势；根据多服务器的特性，探究多服务器异步方式、同步方式与单服务器的关系以及对系统性能的影响。

3.1 实验结果分析

通过固定站点数 $N=5$，服务时间 $\beta=2$，转换时间 $\gamma=1$，比较不同服务器个数的平均排队队长、平均循环周期和平均时延的变化。

图 6－图 8 分别是不同服务器数量下平均排队队长、周期、时延随到达率的变化对比图。由图 6－图 8 可以发现，随着服务器数量的增加，多服务器门限服务系统的队长、周期和时延都逐渐下降。这表明，在相同到达率时，随着服务器数量的增加，系统性能和服务效率也逐渐提高，网络时延变小。

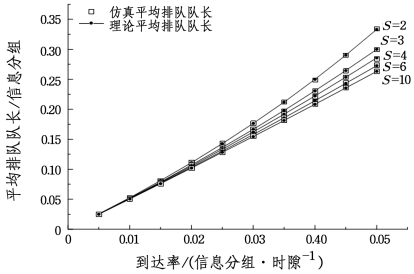

图 6 不同服务器数量平均队长随到达率的变化

Fig. 6 Variation of average length with arrival rate for different number of servers

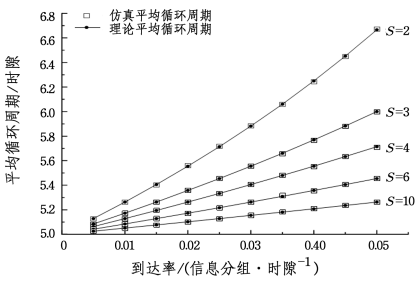

图 7 不同服务器数量平均周期随到达率的变化

Fig. 7 Variation of average period with arrival rate for different number of servers

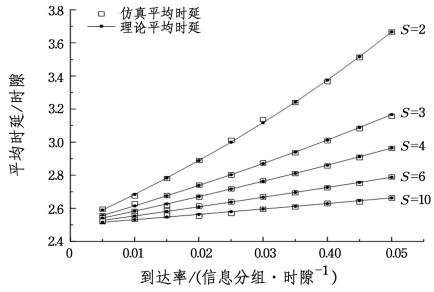

图 8 不同服务器数量平均时延随到达率的变化

Fig. 8 Variation of average delay with arrival rate for different number of servers

进一步分析图 6 可以得出，当到达率较小时，平均队长基本一致，这表明服务器数量的增加对平均排队队长的影响较小。业务量较高，即到达率较高时，平均排队队长有了明显的差距，表明在业务量较高的情况下，使用多服务器能够有效减小平均排队队长。进而得出：在业务量较小时，只需要单个服务器就能满足系统的要求，但在业务量较高的情况下，只使用单服务器是远远不够的，需要增加服务器个数来降低平均排队队长。

图 7 是不同服务器个数下平均循环周期随到达率变化的关系图。在研究轮询系统时，可以采用平均循环周期作为评估标准，即在相同业务量时（到达率相同）平均循环周期越小，系统的服务效率越高。由图 7 可以看出，系统的理论值与仿真值基本一致，很好地验证了理论分析的准确性。随着到达率的提高，平均循环周期逐渐升高；在业务量相同时，增加服务器个数，系统的平均循环周期逐渐下降，这说明系统随着服务器数量的增加越来越稳定。

图 8 给出了不同服务器数量下时延随到达率的变化情况。当保持 $\beta=2$，$\gamma=1$，$N=5$ 不变时，平均时延随着到达率的提高而升高，随着服务器数量的增加，平均时延不断降低。服务器数量越多，平均时延越趋近于平稳。明显可以看出，当服务器数量大于 6 时，平均时延越来越平缓，基本趋于一条直线。这表明到达率较高与到达率较低时平均时延相差不大，基本一致。因此，在某些特殊的情况下，盲目地增加服务器数量是没有效果的，反而会浪费资源。

图 9、图 10 描述了多服务器门限、完全和限定这 3 种服务策略的平均排队队长和平均时延从到达率 0.005 到 0.05 的变化趋势。

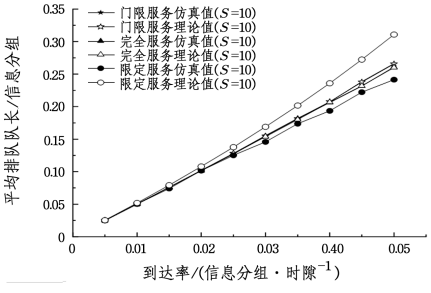

图 9 3 种服务策略平均排队队长的对比图

Fig. 9 Comparison of the average queue length of three service policies

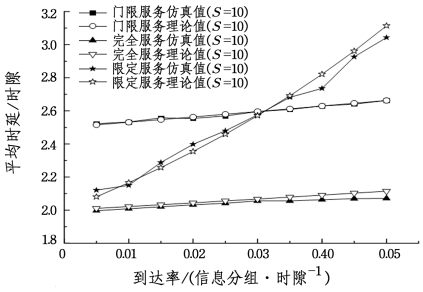

图 10　3 种服务策略平均时延对比图

Fig. 10　Comparison of average delay of three service policies

由图 9、图 10 可知，服务器数量为 10 且到达率大于 0.02 时，限定服务的平均排队队长的理论值与仿真值之间存在较大的差异。完全服务和门限服务的平均排队队长基本保持一致。对于平均时延来说，限定服务与完全服务在小幅度的波动，门限服务相比其他两种服务而言，基本重合，说明了门限服务的稳定性。观察图 9 可知，限定服务的平均排队队长大于门限服务，门限服务大于完全服务。由图 10 可知，当到达率较低时，多服务器门限服务的平均时延大于多服务器限定服务；当到达率较大，超过 0.03 时，门限服务的平均时延小于限定服务，并且随着到达率的提高，限定服务与门限服务之间的差距越来越大。在相同到达率条件下，完全服务的平均时延始终低于门限服务和限定服务。这表明完全服务具有更小的平均时延，系统的性能更好。综上所述，在完全服务系统中，业务量较大时，完全服务的公平性较差，不能保证服务质量，因此为了保证公平性和低时延，门限服务系统更合适。

3.2　同步、异步方式的对比结果分析

多服务器调度时可采用同步方式和异步方式两种。其中同步方式指当站点 i 发送服务请求时，S 个服务器同时为站点 i 服务，直到 S 个服务器都服务结束，S 个服务器才进入下一个发送请求的站点 $i+1$ 进行服务。异步方式指当站点 $i-1$ 和站点 i 都发送服务请求时，由服务器 $j-1$ 和服务器 j 分别为其提供服务。服务器 $j+1$ 直接询问下一个站点。假设站点数为 N，服务器数为 S，若 $N<S$，当 N 个站点全部发送请求时，前 N 个服务器均可为 N 个站点提供服务；若 $N>S$，当 N 个站点全部需要服务时，S 个服务器不够为站点提供服务，故按先请求先服务的原则，为先请求的站点提供服务，待服务器结束服务后，进入后面发送请求的站点进行服务。

图 11 和图 12 是 3 种轮询机制平均排队队长和平均时延的对比图。从图中可以看出，单服务器的队长和时延始终高于多服务器的队长和时延。到达率的提高对多服务器模型的影响较小。在业务量较小时，单服务器与多服务器同步方式相比性能效率提升不大，纵观多服务器异步方式，它在整个业务量范围内均优于其他两种机制。故使用多服务器异步方式能够有效地提高服务效率和系统性能。因此，单服务器与多服务器同步、异步的关系如下：

平均排队队长：单服务器＞同步＞异步

平均时延：单服务器＞同步＞异步

系统性能(服务效率)：单服务器＜同步＜异步

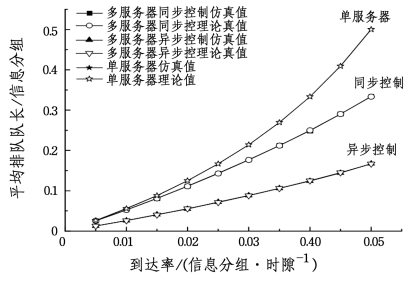

图 11　3 种轮询机制平均排队队长对比图

Fig. 11　Comparison of average queue length of three polling mechanisms

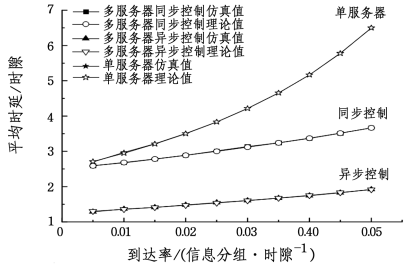

图 12　3 种轮询机制平均时延对比图

Fig. 12　Comparison of average delay of three polling mechanisms

4　基于 BiLSTM 神经网络的性能预测

随着密集站点的接入、大量数据的获取和服务器数量的增加，网络部署和维护难度日益加大。因此本文选取神经网络对轮询系统进行预测，分析系统的性能，降低网络部署和维护的难度。上文提出了一种多服务器门限服务轮询系统模型，并通过仿真实验进行了分析。接下来将构建 BiLSTM 神经网络模型预测其队长和性能，以进一步验证方案的可行性，利用已知数据预测未知数据，探索其趋势。

4.1　BiLSTM 神经网络

针对时间序列数据，传统神经网络难以利用前面数据对后面数据进行预测，而循环神经网络(RNN)由于自身的特性，能够有效保留前一步数据的信息，不停地将数据进行循环操作，是一种对时间序列数据有较好处理效果的神经网络，但存在长期依赖、梯度消失、梯度爆炸等问题[21]。为了避免此类问题的发生，LSTM 神经网络模型被提出，其模型结构如图 13 所示。

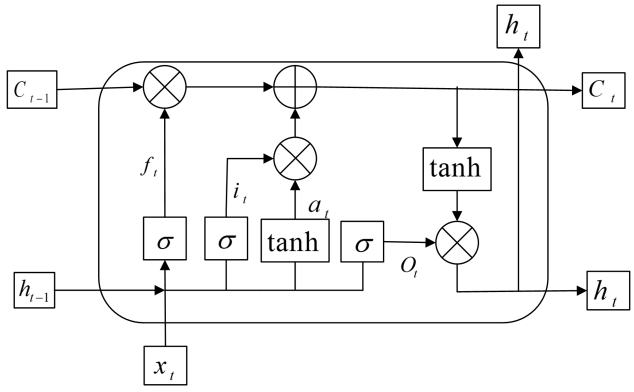

图 13　LSTM 的模型结构

Fig. 13　Model structure of LSTM

LSTM 的核心是细胞状态，即顶部的信息传送带，使信息通过整个细胞单元时基本保持不变。LSTM 基础构成单位是一个记忆存储单元，主要包含 1 个记忆单元模块和 3 个门，即遗忘门、输入门、输出门。它的工作流程如下。

步骤 1　确定细胞状态中需要丢弃的信息。遗忘门通过 Sigmoid 激活函数控制通过的信息，利用 t 时刻的输入 x_t 和 $t-1$ 时刻的隐含层状态 h_{t-1}，得到 0～1 之间的 f_t 值。其中，0 表示完全遗忘，1 表示完全不变。最后根据 f_t 的值决定 C_{t-1} 中的信息，计算式为：

$$f_t=\sigma(w_f h_{t-1}+u_f x_t+b_f) \tag{14}$$

其中，σ 为 Sigmoid 激活函数，w_f 和 u_f 为权重系数，b_f 为遗忘门偏置值。

步骤 2　确定存放在细胞状态中的新信息。首先由输入门决定更新的值；然后利用 tanh 函数创建新的候选值 a_t，计算式为：

$$i_t=\sigma(w_i h_{t-1}+u_i x_t+b_i) \tag{15}$$

$$a_t=\tanh(w_a h_{t-1}+u_a x_t+b_a) \tag{16}$$

其中，i_t 为输入门的值，a_t 为临时细胞状态，w_i，w_a，u_i 和 u_a 为权重系数，b_i 和 b_a 为输入门偏置值。

步骤 3　确定更新旧细胞状态。更新旧细胞状态指把 $t-1$ 时刻的细胞状态更新为当前时刻的细胞状态。利用 $t-1$ 时刻的细胞状态经过遗忘门确定丢弃的信息，再加上通过输入门生成的候选值，得到更新的细胞状态。即当前时刻信息等于之前时刻提取的信息与当前时刻提取的信息之和。具体计算式为：

$$C_t=C_{t-1}\otimes f_t+a_t\otimes i_t \tag{17}$$

其中，C_t 为当前时刻的细胞状态，$\otimes$ 为向量乘积。

步骤 4　确定输出值。该步骤分为两部分，首先，根据激活函数确定细胞状态的输出值，将当前细胞状态通过 tanh 函数得到 -1～1 之间的值；然后，将它与输出值相乘。具体计算式为：

$$O_t=\sigma(w_o h_{t-1}+u_o x_t+b_o) \tag{18}$$

$$h_t=O_t\otimes\tanh(C_t) \tag{19}$$

其中，O_t 为输出门的值，h_t 为当前时刻的输出值，w_o 和 u_o 为权重系数，b_o 为输出门偏置值。

目前，对轮询系统神经网络的研究，大多都是按照时间序列从前往后进行训练，极大降低了数据的利用率，对数据特征获取不充分，并且对时间序列进行预测时，考虑序列的正反信息规律，可以在很大程度上提高预测准确度。而本文利用 BiLSTM 神经网络进行预测，可以更好地获取数据特征，提高预测准确度。BiLSTM 神经网络是由正向和逆向两个 LSTM 构成，相比典型的 LSTM 模型，不仅有从前往后的数据变化，还增加了从后往前的数据变化规律，展现出了该模型的优越性。BiLSTM 模型的结构如图 14 所示。

BiLSTM 神经网络的工作流程如下。

步骤 1　确定模型自前向后（正向）的输出。对前向层某一时刻 t_i（$i=1,2,\cdots,n$）进行正向计算，获取隐含层输出 M_b，计算式为：

$$M_b=f(w_1 x_t+w_2 M_{b-1}) \tag{20}$$

步骤 2　确定模型自后往前（逆向）的输出。对后向层 t_i（$i=n,n-1,\cdots,2,1$）时刻进行反向计算，获取隐含层输出 M_i，计算式为：

$$M_i=f(w_3 x_t+w_5 M_{i-1}) \tag{21}$$

步骤 3　确定 t_i 时刻的输出值 y_t。最终输出值由前向层和后向层组成。计算式为：

$$y_t=g(w_4 M_b+w_6 M_i) \tag{22}$$

其中，w_i（$i=1,2,\cdots,6$）为权重系数。

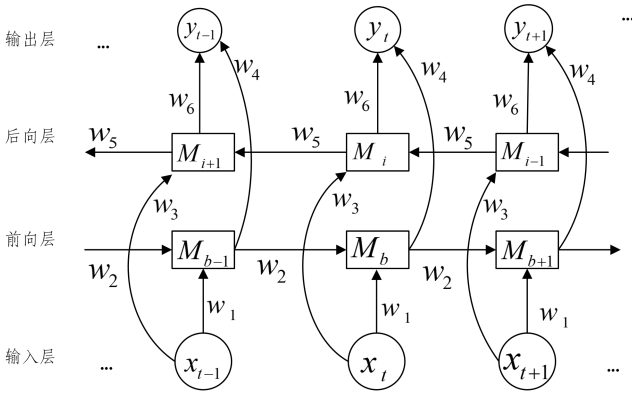

图 14　BiLSTM 的模型结构

Fig. 14　Model structure of BiLSTM

4.2　数据准备及模型搭建

平均队长和时延是衡量轮询系统性能的指标，平均队长和时延越小，性能就越好。为了实现神经网络预测，在上述实验的基础上，扩大实验规模，获取 500 组不同到达率下的平均队长数据，将到达率与平均队长的关系抽象为一组时间序列问题。将数据的 70% 作为训练集，将数据的 30% 为测试集，并且将数据归一化到[0～1]之间，计算式为：

$$y_{gyh}=\frac{y-y_{\min}}{y_{\max}-y_{\min}} \tag{23}$$

其中，y_{gyh} 表示归一化后的值，y 表示实际值，$y_{\max}$ 表示数据中的最大值，$y_{\min}$ 表示数据中的最小值。

获取数据集后，在 python 环境中使用 tensorflow 框架搭建 BiLSTM 神经网络模型。为了使预测效果最佳，经过反复实验，设定网络隐含神经元的数量为 32，训练次数为 200，学习率为 0.001，误差函数使用均方误差，计算式为：

$$MSE=\frac{1}{N}\sum_{i=1}^{N}(y_{\text{true}}-y_{\text{pred}})^2 \tag{24}$$

其中，N 表示数据集的总数量，y_{true} 表示真实值，y_{pred} 表示预测值。

5　预测分析

将数据导入 BiLSTM 神经网络中，利用训练好的模型进行预测。本次实验使用 500 组数据，其预测结果如下。

图 15 给出了不同服务器数量下的平均队长预测变化。可以发现，预测值与理论值基本一致，误差保持在较小的范围内，预测结果准确，随着服务器数量的增加，平均队长逐渐减小。这表明该模型能够准确预测多服务器门限服务系统的性能。从图 16 可以看出，该模型在任何机制下都可较好地进行预测。其中，针对平均队长，异步控制优于同步控制，同步控

制优于单服务器，与上述仿真实验分析的结果一致。

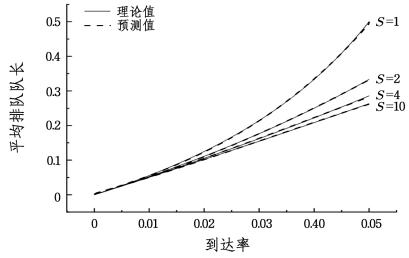

图 15 不同服务器数量平均队长的预测变化

Fig. 15 Predicted change in average length for different server numbers

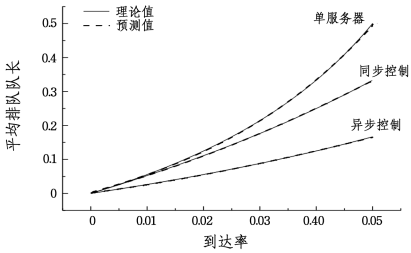

图 16 3 种轮询机制平均队长预测变化

Fig. 16 Predicted change in average length for three polling mechanisms

上述实验均是利用前三步预测下一步，其结果较好。在实际通信中，轮询系统模型有时较为复杂，计算难度较大，使用神经网络模型对未知数据的预测能够降低计算难度，有助于分析轮询系统的运行状态，进行合理的性能评估。图 17 是未知数据的预测变化图，利用滑动窗口方式将已知数据输入，以预测下一个未知数据，再依次将未知数据预测值加入到已知数据中，预测再下一个未知数据。图 17 中，预测了服务器数量为 1，2，4，10 的平均队长。可以得出，未知预测值曲线与理论值曲线的趋势相同，随着到达率的提高，平均队长增加。通过进一步观察可知，整体的预测误差保持在可控范围内，但随着到达率接近 0.06，队长的预测误差逐渐变大，这表明未知数据预测得越多，误差累积就越大。这是因为在后期数据的预测过程中，把前期输出的预测值作为输入。

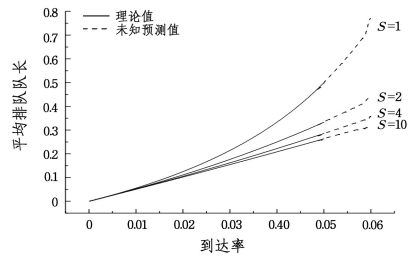

图 17 不同服务器数量队长的未知预测变化

Fig. 17 Unknown predicted variation in the average length for different server numbers

为了更进一步验证 BiLSTM 模型的准确性，本文利用 LSTM 模型对同一个 500 组实验数据进行预测，其预测结果如图 18 所示。从图中可以看出，BiLSTM 的预测效果优于 LSTM。其原因是，BiLSTM 模型是由正向和反向两个 LSTM 单元组成，能够分析从前往后和从后往前的数据特征，能更加充分地挖掘数据信息，故预测效果较好。实验显示，BiLSTM 的预测误差是 0.000 8，而 LSTM 的预测误差是 0.002，从预测误差的角度也可得出 BiLSTM 的预测效果优于 LSTM。

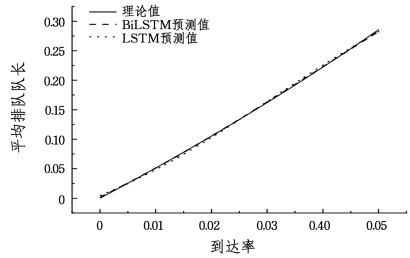

图 18 不同神经网络平均队长的预测变化

Fig. 18 Predicted change in average length for different neural networks

结束语 本文采用理论分析和仿真实验对多服务器门限服务系统进行分析，比较了平均队长、周期和时延 3 个指标。结果表明，多服务器接入方式可以很好地改善系统效率和性能，降低网络时延。相比单服务器接入方式，多服务器接入方式在业务量大、密集站点接入的情况下更适用。其次，对比多服务器门限、完全、限定 3 种服务策略，综合考虑公平性与低时延性，多服务器门限服务系统更好。同时，将多服务器门限服务系统的同步和异步两种方式进行比较，得出异步方式更好。最后，构建 BiLSTM 模型来预测轮询系统，预测效果较好，降低了计算难度。最终实验结果表明，所提系统方案是可行的，具有指导意义。

本文仅仅是对多服务器门限服务系统进行考虑，在后期工作中，首先将多服务器接入方式应用到两种不同的服务策略上，即探究多服务器两级轮询系统，然后把多服务器接入方式运用到非对称系统模型中，探究其特性。

参 考 文 献

[1] TRAN T X，POMPILI D. Joint Task Offloading and Resource Allocation for Multi-Server Mobile Edge Computing Networks[C]// IEEE Transactions on Vehicular Technology. 2019：856-868.

[2] TEYMOORI P，SOHRABY K，KIM K. A Fair and Efficient Resource Allocation Scheme for Multi-Server Distributed Systems and Networks[C]// IEEE Transactions on Mobile Computing. 2016：2137-2150.

[3] BOON M A A，WINANDS E M M，ADAN I J B F，et al. Closed-form waiting time approximations for polling systems[J]. Performance Evaluation，2011，68(3)：290-306.

[4] LI H F，LING Z G，DING H W，et al. Mixed Service Policy Polling Feature Analysis and Access Control System for Energy Saving Design[J]. Journal of Internet Technology，2019，9(6)：37-40，43.

[5] BAO L Y，ZHAO D F，DING H W. Research on Load Balance Strategy of the Double Server in the Synchronous Dispatch Mecha-nism of Polling[J]. Journal of Yunnan University(Natu-

ral Sciences Edition), 2009, 31(s1): 1-4, 8.

[6] BAO L Y, ZHAO D F, DING H W. An Approximate Algorithm Analysis of Dual-server Asynchronous Control Policy Polling System Performance[C]// Proceedings of 2009 China University Communication Academic Conference. 2009: 401-405.

[7] HUANG W H, MA Z, DAI X F, et al. Fuzzy Clustering Based Load Balancing Algorithm with Feature Weighted[J]. Journal of Xidian University, 2017, 44(2): 127-132.

[8] TAO X L, WEI Y, WANG Y. A Load Balancing Method Based on Hierarchy and Multi-agent for Cloud Computing Platform [J]. Acta Electronica Sinica, 2016, 44(9): 2106-2113.

[9] ZHAO Z H, LI Q, YANG S P, et al. Based on BiLSTM and Attention Mechanism of Residual Service Life Prediction Research [J]. Journal of Vibration and Shock, 2022, 41(6): 44-50, 196.

[10] ZHU L J, XUN Z H, WANG Y X, et al. Short-term Power Load Forecasting Based on CNN-BiLSTM[J]. Power Grid Technology, 2021, 45(11): 4532-4539.

[11] LIU Q L, ZHAO D F, DING H W, et al. The Evolution and Development of Polling System [J]. Wireless Communication Technology, 2013, 39(2): 55-59.

[12] YANG Z J, ZHENG H Y, DING H W. Wireless LAN Multi-robot System Polling MAC Protocol Study[J]. Computer Application Researh, 2022, 39(4)6: 1178-1182.

[13] CHEN C L. Research on MAC enhancement technology of low delay and high reliability wireless LAN[D]. Xi'an University of Electronic Science and Technology, 2019.

[14] EDIRISINGHE S, RANAWEERA C, LIM C, et al. Universal optical network architecture for future wireless LANs[J]. Journal of Optical Communications and Networking, 2021, 13(9): D93-D102.

[15] HAYAKAWA M, ITO Y. Study on an Appropriate Value of the Maximum Transmission Rate over Congested IEEE802. 11g Wireless LAN Based on Web QoE[C]// 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE). 2019.

[16] YANG Z J, MAO L, DING H W. Analysis of Dual-server Polling Access Control Protocol[C]// 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference(ITNEC), 2020: 2607-2612.

[17] CHEN B, WU L, ZEADALLY S, et al. Dual-server Public-Key Authenticated Encryption with Keyword Search [C] // IEEE Transactions on Cloud Computing. 2022: 322-333.

[18] YANG Z J, MAO L DING H W, et al. Analysis of Continuous-time Two-level Complete Polling Access MAC Protocol [J]. Computer Engineering and Applications, 2022, 58(9): 136-143.

[19] YANG Z J, MAO L, YAN B, et al. Performance analysis and prediction of asymmetric two-level priority polling system based on BP neural network [J]. Applied Soft Computing, 2021, 29(6): 1046-1053.

[20] YANG Z J, LIU Z, DING H W. Research on Performance of Continuous-time Exhaustive Service and Gated Service Two-level Polling System[J]. Journal of Computer Applications, 2019, 39(7): 2019-2023.

[21] XIE X Y, ZHOU J H, ZHANG Y J, et al. An Ultra-short-term Power Generate on Forecasting Method Based on W-BiLSTM [J]. Automation of Electric Power Systems, 2021, 45(8): 175-184.

YANG Zhijun, born in 1968, Ph.D, researcher. His main research interests include computer network and communication, polling control system, and deep learning.

(责任编辑:喻藜)