5G网络切片场景中基于预测的虚拟网络功能动态部署算法

唐 伦 周 钰* 杨友超 赵国繁 陈前斌
(重庆邮电大学通信与信息工程学院 重庆 400065)
(重庆邮电大学移动通信重点实验室 重庆 400065)

摘 要：针对无线虚拟化网络在时间域上业务请求的动态变化和信息反馈时延导致虚拟资源分配的不合理，该文提出一种基于长短时记忆(LSTM)网络的流量感知算法，该算法通过服务功能链(SFC)的历史队列信息来预测未来负载状态。基于预测的结果，联合考虑虚拟网络功能(VNF)的调度问题和相应的计算资源分配问题，提出一种基于最大最小蚁群算法(MMACA)的虚拟网络功能动态部署方法，在满足未来队列不溢出的最低资源需求的前提下，采用按需分配的方式最大化计算资源利用率。仿真结果表明，该文提出的基于LSTM神经网络预测模型能够获得很好的预测效果，实现了网络的在线监测；基于MMACA的VNF部署方法有效降低了比特丢失率的同时也降低了整体VNF调度产生的平均端到端时延。

关键词：5G网络切片；资源分配；流量感知；预测；虚拟网络功能F调度
中图分类号：TN929.5 **文献标识码**：A **文章编号**：1009-5896(2019)09-2071-08
DOI: 10.11999/JEIT180894

Virtual Network Function Dynamic Deployment Algorithm Based on Prediction for 5G Network Slicing

TANG Lun ZHOU Yu YANG Youchao ZHAO Guofan CHEN Qianbin
(*School of Communication and Information Engineering, Chongqing University of Post and Telecommunications, Chongqing* 400065, *China*)
(*Key Laboratory of Mobile Communication Technology, Chongqing University of Post and Telecommunications, Chongqing* 400065, *China*)

Abstract: In order to solve the unreasonable virtual resource allocation caused by the dynamic change of service request and delay of information feedback in wireless virtualized network, a traffic-aware algorithm which exploits historical Service Function Chaining (SFC) queue information to predict future load state based on Long Short-Term Memory (LSTM) network is proposed. With the prediction results, the Virtual Network Function (VNF) deployment and the corresponding computing resource allocation problems are studied, and a VNFs' deployment method based on Maximum and Minimum Ant Colony Algorithm (MMACA) is developed. On the premise of satisfying the minimum resource demand for future queue non-overflow, the on-demand allocation method is used to maximize the computing resource utilization. Simulation results show that the prediction model based on LSTM neural network in this paper obtains good prediction results and realizes online monitoring of the network. The Maximum and Minimum Ant Colony Algorithm based VNF deployment method reduces effectively the bit loss rate and the average end-to-end delay caused by overall VNFs' scheduling at the same time.

Key words: 5G network slicing; Resource allocation; Traffic-aware; Prediction; Virtual Network Function (VNF) scheduling

收稿日期：2018-09-18；改回日期：2019-02-20；网络出版：2019-03-21
*通信作者： 周钰 137068966@qq.com
基金项目：国家自然科学基金(61571073)
Foundation Item: The National Natural Science Foundation of China (61571073)

1 引言

5G网络架构下涌现的多种应用场景在移动性、安全性、时延和可靠性等方面要求各不相同[1]。下一代移动网络联盟提出了网络切片的概念[2]，在该场景下的服务功能链(Service Function Chaining, SFC)创造了动态的网络服务，由于SFC中的虚拟网络功能(Virtual Network Function, VNF)运行在通用服务器中，这种业务处理架构保证了VNF的灵活性和可调整性[3]。然而，如何设计有效的VNF部署、网络资源配置方案是一个巨大的挑战[4]。

文献[5]采用遗传算法解决了服务功能链的部署问题，虽然实现了各条链间更公平的资源分配，但只能解决单一调度周期上的资源调度问题。文献[6]设计了一种基于马尔可夫近似算法最大化混合型SFC网络效用，但并未考虑网络负载的动态变化对业务速率需求的影响。文献[7]基于无线虚拟化网络内的缓存空间和无线带宽调度问题建立了一种博弈机制，但并未考虑虚拟网络间的差异化需求。文献[8]提出了一种基于深度学习的虚拟机工作负载预测方法，然而该模型仅适用于单个虚拟机的资源预测。文献[9]提出一种基于图形神经网络(GNN)的算法预测VNF对资源的需求，并通过Openstack平台实现了VNF的调度和动态的资源管理，但该方法过于依赖网络的拓扑信息。

综上所述，在研究SFC部署的相关文献中，大多数工作只停留在解决单一部署周期的资源调度问题上，而在实际网络场景中，若未动态地分配网络资源处理变化的业务请求，可能会引起数据的积压而导致队列溢出概率增大、端到端时延增加等问题，因此网络应根据当前的队列状态动态地调整资源分配以提供稳定的服务[10]；同时，将预测与SFC的部署相结合的研究较少，导致在VNF的调度过程由于滞后性引起服务性能的下降。本文基于业务请求动态变化的5G网络切片应用场景，设计了一种基于长短时记忆网络(Long Short-Term Memory, LSTM)预测的虚拟网络功能部署算法，本文的主要贡献总结如下：(1)设计基于时延感知的SFC队列缓存模型，并在包含多种时间尺度的时间域上分配虚拟资源；(2)考虑到用户请求业务的动态性和信息反馈的延迟引起不合理的虚拟网络资源分配或缓存溢出概率增大等问题，提出一种基于LSTM神经网络的流量感知算法，通过各个切片业务的历史队列信息来预测未来的负载情况，从而实现了网络的在线监测；(3)基于对多个切片业务负载情况的预测结果，提出了一种动态的VNF调度与资源分配方案，引入了一种最大最小蚁群算法(Maximum and Minimum Ant Colony Algorithm, MMACA)实现多条服务功能链的动态部署，根据各个切片的队列大小实时地为其分配虚拟网络资源。

2 系统模型

在5G网络功能虚拟化、无线云化环境中，资源的分配是在“池”的层面上进行的，这种方式带来了资源的灵活分配，从而实现了更加动态的网络服务，满足更多样化的需求[11]。

图1为网络切片场景下的SFC部署系统架构图，从左至右依次为无线虚拟网络用户、虚拟网络

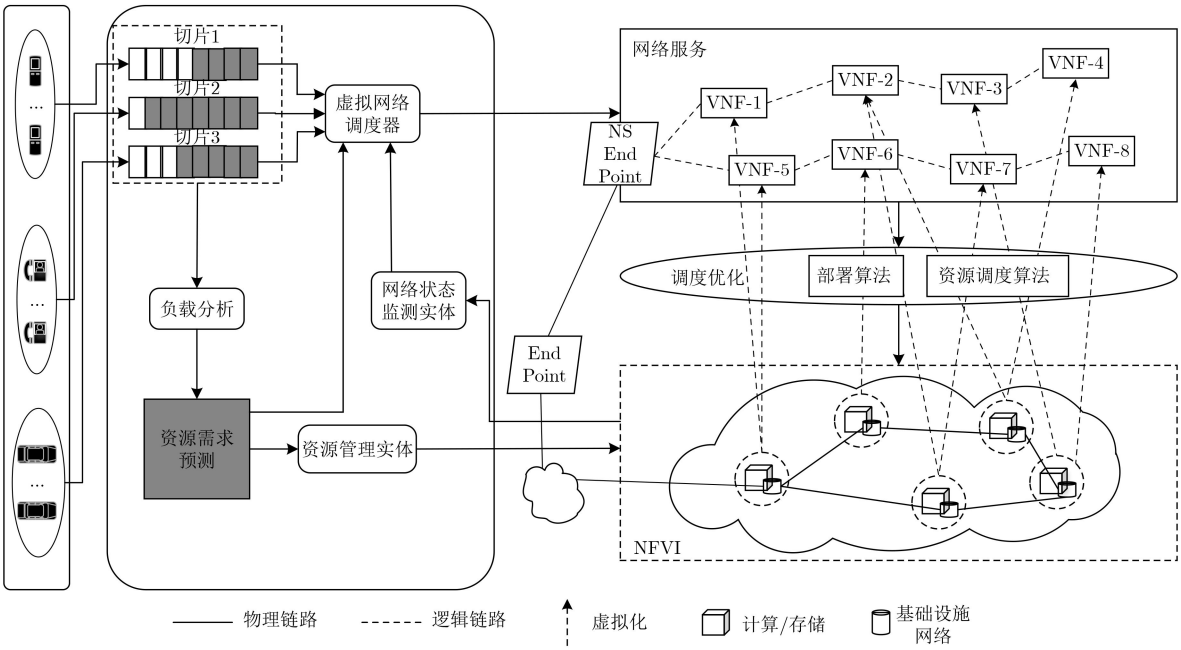

图 1 网络切片场景下的SFC部署系统架构图

管理平台、虚拟网络功能调度层、虚拟资源池。在该系统中，虚拟资源池中的云服务器提供包含计算资源、缓存资源、带宽资源等多种类型的虚拟网络资源，虚拟网络管理平台则根据虚拟网络用户的业务状态，QoS需求等实现VNF的调度、虚拟网络资源的灵活分配。本文设计的虚拟网络管理平台由业务请求单元、负载分析模块、资源管理实体、网络状态监控实体、虚拟网络调度器等部分组成。其中，业务请求单元用于缓存各切片用户新到达的数据包；负载分析模块用于分析各切片的缓存负载特征，并预测下一周期负载状态；虚拟网络调度器则根据负载分析模块的评估结果决定每条SFC的部署方案；资源管理实体在SFC完成部署后分配各个虚拟网络功能模块获得的最佳虚拟资源量；网络状态监控实体的作用是观察各虚拟资源的实时状态。

3　网络模型

3.1　基于时延的服务功能链队列模型

本文将不同用户对同一类型网络业务请求的数据集合定义为切片，用S表示。F表示所有VNF模块的集合，对于第i个网络切片S_i，业务请求的VNF集合用$F_i=\{f_{i1},f_{i2},\cdots,f_{ij},\cdots,f_{iJ}\}$表示。其中，$i\in[1,|S|]_z, j\in[1,|F_i|]_z$，$f_{ij}$表示网络切片$i$需要调度的第$j$个VNF。虚拟网络拓扑由无向图$G=(V,E)$表示，其中，$V$表示虚拟节点的集合，$E$表示虚拟链路的集合。$F'_m(m\in[1,|V|]_z)$表示虚拟节点$m$能处理的VNF种类集合，例如$F'_m=\{f_1,f_3,f_6\}$表示虚拟节点$m$能够处理虚拟网络功能$f_1$, f_3, f_6，本文假设网络内的虚拟链路带宽资源充足。令$N_{f_{ij}}(i\in[1,|S|]_z, j\in[1,|F_i|]_z)\subseteq V$表示能够执行虚拟化网络功能$f_{ij}$的虚拟节点集合。

$x^m_{f_{ij}}$表示网络切片i需要调度的第j个VNF是否选择在虚拟节点m上执行

$$x^m_{f_{ij}}=\begin{cases}1, & m\text{执行}f_{ij},\text{其中}m\in N_{f_{ij}}\\ 0, & \text{其它}\end{cases}\tag{1}$$

$y^l_{f_{ij}}$表示执行完虚拟网络功能f_{ij}后传输数据流的虚拟链路的选择

$$y^l_{f_{ij}}=\begin{cases}x^m_{f_{ij}}, & m\text{执行}f_{ij},\text{选择第}l\text{条虚拟链路}\\ & \text{传输数据流},\text{其中}l\in[1,|E_m|_z]\\ 0, & \text{其它}\end{cases}\tag{2}$$

本文将网络运行的时间维度分为若干个时隙，用$\Gamma=\{1,2,\cdots,t,\cdots,T\}$表示网络运行的时隙集合，其中，定义每个时隙$t$的持续时间为$T_s$。因此在时隙$t$内，$R_i(t)$表示在时隙$t$内提供给切片$S_i$的服务速率。令$C_{f_{ij}}(t)=\eta_{f_{ij}}\cdot R_i(t)$，其中，$C_{f_{ij}}(t)$表示切片$S_i$在时隙$t$内节点执行$f_{ij}$提供一定服务速率所需的计算资源。$\eta$表示服务速率系数，该系数表示不同种类VNF提供一定服务速率所对应的资源关系[12]。

假设$Q_i(t)$表示在时隙t切片S_i的队列长度，$A_i(t)$表示数据包的到达过程且服从参数为λ_i的泊松分布，所有用户的包到达过程在不同调度时隙是独立分布的，令$P_i(t)$代表数据包大小，数据包大小服从均值为$\bar{P}_i$的指数分布[13]。那么平均包的处理速率为$\mu_i=R_i/\bar{P}_i$。所以缓存中队列的长度更新过程表示为

$$Q_i(t+1)=\max[Q_i(t)-D_i(t),0]+A_i(t)\tag{3}$$

其中，$D_i(t)=\mu_i(t)\cdot T_s$表示在时隙t内被处理的数据包数目。

令$Q_i^{\max}$表示第i个切片队列所允许的最大缓存大小，为了降低切片的平均比特丢失率，计算防止切片队列溢出所需最低的服务速率，对于任意t内，第i个切片队列长度的增量表示为$I_i(t)=A_i(t)-D_i(t)$，且切片队列不溢出需满足$Q_i(t+1)-Q_i^{\max}\le 0$，因此服务速率应保证

$$R_i(t)\ge(Q_i(t)+A_i(t)-Q_i^{\max})/T_s\tag{4}$$

本文考虑的端到端时延主要由排队时延和处理时延所决定。由于假设虚拟链路带宽资源是足够的，因此假设传输时延、VNF迁移产生的时延以及数据包大小差异造成的等待时延忽略不计。令$\overline{T}^i_w$, $\overline{T}^i_p$分别表示切片S_i数据包平均排队时延、在整个网络相应虚拟节点上的平均处理时延。因此端到端时延τ可表示为

$$\tau_i=\overline{T}^i_p+\overline{T}^i_w\tag{5}$$

由Little定理可以得到切片i的平均排队时延与平均队列长度的关系为$\bar{Q}_i=\lambda_i\overline{T}^i_w$，处理时延$X_i$由执行各个VNF的时延$X_{f_{ij}}$组成，且$X_{f_{ij}}=P_i/R_i$，$X_i=X_{f_{i1}}+X_{f_{ij}}+\cdots+X_{f_{iJ}}$，数据包大小相互独立，同时各个节点服务速率保持一致性，所以传输切片S_i的数据包总平均端到端时延为

$$\begin{aligned}\tau_i&=\overline{T}^i_p+\overline{T}^i_w\\&=\lim_{T\to\infty}\frac{1}{T}\mathrm{E}\left\{\sum_{j=0}^{J}\sum_{t=0}^{T-1}\frac{\bar{P}_i}{R_i(t)}+\sum_{t=0}^{T-1}\frac{Q_i(t)}{\lambda_i}\right\}\end{aligned}\tag{6}$$

本文的目标是从而最小化网络内多条服务功能链VNF整体调度产生的端到端时延，因此SFC动态部署问题可建立成式(7)的数学模型

$$\min\tau,\ \tau=\max\{\tau_1,\tau_2,\cdots,\tau_i\}\tag{7}$$

式(8)中相应约束说明：C1保证了处理切片S_i的1个VNF只能在1个相应的虚拟节点上运行。C2保证了执行完f_{ij}后传输数据流的虚拟链路不超过

1条。C3保证了对于f_{ij}在虚拟节点上执行后数据通过链路l传输，如果该链路的头部节点是m，则表明执行f_{ij}的虚拟节点一定是m。C4保证了数据流从第$j-1$个VNF到第j个VNF只使用了1条链路l。C5保证了虚拟节点m上的虚拟化网络功能f'同时只能被1个切片的服务功能链所调度。C6确保VNFs的资源需求不能超过虚拟节点的资源限制。

$$
\begin{aligned}
&\text{s.t.}\\
&\text{C1}: \sum_{m\in N_{f_{ij}}} x_{f_{ij}}^{m}=1, i\in[1,|S|]_z, j\in[1,|F_i|]_z;\\
&\text{C2}: \sum_{l\in[1,E_m]_z} y_{f_{ij}}^{l} x_{f_{ij}}^{m}\le 1, i\in[1,|S|]_z,\\
&\quad j\in[1,|F_i|]_z;\\
&\text{C3}: x_{f_{ij}}^{m}=\sum_{m=l.\text{head}} y_{f_{ij}}^{l}, i\in[1,|S|]_z,\\
&\quad j\in[1,|F_i|-1]_z, m\in N_{f_{ij}};\\
&\text{C4}: \sum_{m=n.\text{tail}} y_{f_{ij-1}}^{l}=x_{f_{ij}}^{m}, i\in[1,|S|]_z,\\
&\quad j\in[2,|F_i|]_z, m\in N_{f_{ij}};\\
&\text{C5}: \sum_{i\in[1,|S|]_z} z_{f'}^{i}=1, f'\in F_m';\\
&\text{C6}: \sum_{i}^{I}\sum_{j}^{J} C_{f_{ij}}\le \sum_{m}^{M} C_{\text{total}}(m)
\end{aligned}
\tag{8}
$$

3.2 基于LSTM神经网络的预测方法

由上一节可知各切片业务的负载特征直接影响着SFC部署的有效性，而采用预测的机制实现网络的在线监测能解决资源管理的滞后性，因此本文采用基于LSTM的预测方法，提前预测缓存内的未来负载状态。该方法定义每一个节点n对应的特征F_n，通过这些特征信息，LSTM模型为每个节点n制定了隐藏层状态h_n，进而决定每个节点的输出o_n。基于LSTM的动态资源管理系统对于单个切片的模型如图2所示

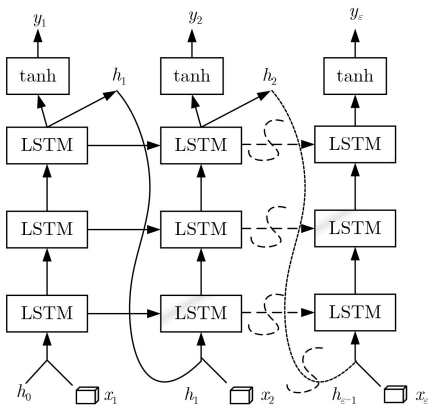

图 2 单个切片基于LSTM的资源预测模型

3.2.1 特征与状态

切片特征由观察到的当前缓存中队列长度Q_i与用户请求该业务的数据包到达率A_i组成，具体为在虚拟网络$G=(V,E)$中，对于服务功能链i第j个VNF，令$C_{f_{ij}}$表示防止队列溢出的最低计算资源需求，为了简化问题，本文仅考虑CPU资源的使用。所以切片i的特征可以表示为$x_i=\{A_i,Q_i\}$，所以对于某一时刻t，神经网络输入为$x_i(t)=\{A_i(t),Q_i(t)\}$。另外定义1个历史数据集的窗口长度ε，在历史时刻$t-\varepsilon$至t的范围内，网络模型输入的数据集可以表示为

$$
\begin{bmatrix} x_i(t)\\ x_i(t-1)\\ \vdots\\ x_i(t-\varepsilon)\end{bmatrix}=\begin{bmatrix} A_i(t) & Q_i(t)\\ A_i(t-1) & Q_i(t-1)\\ \vdots & \vdots\\ A_i(t-\varepsilon) & Q_i(t-\varepsilon)\end{bmatrix} \tag{9}
$$

本文构建的LSTM网络模型可以描述为

$$
\begin{pmatrix} i_t\\ f_t\\ o_t\\ g_t\end{pmatrix}=\begin{pmatrix}\sigma\\ \sigma\\ \sigma\\ \tanh\end{pmatrix}\boldsymbol{\Lambda}\begin{pmatrix} h_{t-1}\\ x_t\end{pmatrix} \tag{10}
$$

$$
c_t=f_t\odot c_{t-1}+i_t\odot g_t \tag{11}
$$

$$
h_t=o_t\odot\tanh(c_t) \tag{12}
$$

其中，i_t为输入门，f_t为遗忘门，o_t为输出门，g_t决定了输入有多少保存到单元状态中，c_t为单元状态，$\boldsymbol{\Lambda}$为包含学习的参数仿射变换矩阵，h_{t-1}为隐藏层，σ和$\odot$分别表示激活函数sigmoid和元素级乘积。

本文的模型将历史的队列状态以及用户请求业务的数据包历史到达率作为输入，输出为缓存内的未来负载状态进而得到服务功能链VNF具体资源需求的预测结果。该过程具体的描述如表1所示。

表 1 基于LSTM的VNF资源需求预测模型

1 初始化：W，迭代次数$\kappa=0$，状态$h(\kappa)=0$
2 *网络状态的监测*
3 监测不同时刻用户请求业务的数据包到达率
4 利用SFC部署算法更新队列状态并记录当前队列大小
5 *状态计算*
6 while ($\kappa<T$) do
7 通过式(10)—式(12)计算$h(\kappa+1)$
8 $\kappa=\kappa+1$
9 end while
10 *输出计算*
11 计算不同切片业务负载的预测值
12 计算输出各个VNF资源需求预测值$C_{f_{ij}}$

3.2.2 网络学习与训练

为了实现VNF资源需求精确的预测，权重函数需要反复的训练，因此针对本文解决的问题，需要大量的样本数据，其中包含切片历史特征信息$x_i(t-\varepsilon), x_i(t-\varepsilon+1), \cdots, x_i(t)$，未来时间段内负载状态的实际值$\xi_t$。训练的目标是式(13)的2次成本惩罚函数最小化

$$G_w = \sum_{i\in I}\left(\frac{1}{2}(o_i-\xi_i)^2+\alpha(o_i-\psi)^2+\gamma\sum_w w^2\right) \tag{13}$$

其中，式(13)的第1项为损失函数，o_i为预测值，ξ_i为真实值；第2、第3项分别为惩罚项，L2正则化约束项，α为惩罚系数，γ为权值衰减系数，ψ为阈值，w表示模型中所有的权重参数。训练的目标是找到最佳的权重W(拟合数据的特征)使得成本函数最小化，具体的学习流程如表2所示。

表 2　学习与训练权重

1 迭代$\kappa=0$时，采用Xavier[14]初始化权重，即令权重的概率分布函数服从$W\sim U\left[-\frac{\sqrt{6}}{\sqrt{\chi_p+\chi_{p+1}}},\frac{\sqrt{6}}{\sqrt{\chi_p+\chi_{p+1}}}\right]$的均匀分布，其中$\chi$为网络层数，$p$为神经元个数
2 while(未达到训练要求标准)do
3　使用表1中算法计算状态$h(\kappa)$和输出$o(\kappa)$等参数
4　利用反向传播算法(BPTT)，使用惩罚函数式(13)计算惩罚函数的梯度
5　通过式(14)更新权重W
6　$\kappa=\kappa+1$
7 end while

$$W(\kappa+1) = W(\kappa) - \varphi\frac{\partial G}{\partial W} \tag{14}$$

式(14)为权重的更新，$\frac{\partial G}{\partial W}$表示惩罚函数的梯度，$\varphi$表示学习率，$G$表示惩罚函数。

3.3 基于最大最小蚁群算法的服务功能链部署方法

在应用蚁群算法求解服务功能链问题的过程中，使用一个2维矩阵pathMatrix表示多条服务功能链的部署列表，其中，2维矩阵的行数为$\sum_{i\in[1,|S|]_z}F_i$，列数为$|M|$。将每只蚂蚁在1次循环内的搜索到的VNF调度路径看成是服务功能链部署的1个解，考虑到每种业务用户请求数据大小的不同，从而导致VNF处理时延的要求有一定的差异，这些因素都将影响算法的寻优过程。具体的算法设计流程如表3所示。

在自适应蚁群算法中，将信息素浓度挥发因子ρ，信息素因子∂，启发函数重要程度因子β分别限制在区间$[\rho_{\min},\rho_{\max}]$, $[\partial_{\min},\partial_{\max}]$, $[\beta_{\min},\beta_{\max}]$，其中$\rho_{\min}$，$\partial_{\min}$，$\beta_{\min}$代表了$\rho$，$\partial$和$\beta$的最小值，$\rho_{\max}$，$\partial_{\max}$，$\beta_{\max}$代表了$\rho$, ∂和β的最大值。因此，当算法陷入局部最优时对以上3个参数按式(15)—式(17)更新。

表 3　基于最大最小蚁群算法的多条服务功能链部署算法

1 初始化：蚂蚁规模antNum、信息素因子∂、启发函数重要程度因子β等参数
2 for itCount = 1 : iteratorNum do
3　for antCount = 1 : antNum do
4　*VNFs的调度*
5　for $i=1:I$ do
6　for $j=1:J$ do
7　根据式(18)计算转移概率$P(k)$并采用轮盘赌法选择下一个VNF部署的节点m
8　更新路径矩阵pathMatrix和节点剩余资源向量$\boldsymbol{C}_{\mathrm{re}}(m)$
9　end
10　end
11　end
12 *虚拟节点计算资源的分配*
13　//对每个虚拟节点上的VNF按需求比例进行资源的分配
14　for $m=1:M$ do
15　$C_{\max}(f_{ij}) = C_{\mathrm{total}}(m)\cdot C_{\mathrm{pr}}(f_{ij})\Big/\sum_{i=1}^{I}\sum_{j=1}^{J}C_{\mathrm{pr}}(f_{ij})\cdot x_{f_{ij}}^{m}$
16　end
17　//计算每条服务功能链提供最大服务速率时实际分配VNF的计算资源
18　for $i=1:I$ do
19　$C_{f_{ij}} = \min_{j\in J}\{C_{\max}(f_{ij})/\eta_{f_{ij}}\}\cdot\eta_{f_{ij}}$
20　end
21 根据式(19)更新信息素矩阵；根据式(15)—式(17)，式(22)对$\rho,\partial,\beta,\zeta$进行更新
22 end
23 多次迭代后得到近似最优的SFC部署方案$\mathrm{Dep}_{\mathrm{SFC}_i}$
24　*计算服务功能链VNF平均调度时延*
25　for $i=1:I$ do
26　根据式(6)计算平均端到端时延τ_i
27　end

$$\rho(\mathrm{iN}+1) = \begin{cases} 0.95\rho(\mathrm{iN}), & 0.95\rho(\mathrm{iN})\geq\rho_{\min} \\ \rho_{\min}, & \text{其它} \end{cases} \tag{15}$$

$$\partial(\mathrm{iN}+1) = \begin{cases} 1.05\partial(\mathrm{iN}), & 1.05\partial(\mathrm{iN})<\partial_{\min} \\ \partial_{\max}, & \text{其它} \end{cases} \tag{16}$$

$$\beta(\mathrm{iN}+1) = \begin{cases} 1.1\beta(\mathrm{iN}), & 1.1\beta(\mathrm{iN})<\beta_{\min} \\ \beta_{\max}, & \text{其它} \end{cases} \tag{17}$$

在进行服务功能链路径选择的过程中，当前节点转移至k节点的状态转移概率为

$$P(k)=\frac{\zeta_{\mathrm{ck}}^{\partial}\cdot\delta_{k}^{\beta}}{\sum\limits_{m\in N_{f_{ij}}}\zeta_{\mathrm{cm}}^{\partial}\cdot\delta_{m}^{\beta}} \tag{18}$$

c表示执行f_{ij-1}的虚拟节点，k表示下一节点，$N_{f_{ij}}$表示能够执行f_{ij}的虚拟节点集合，$k\in N_{f_{ij}}$，∂表示信息素因子，β为启发函数重要程度因子，δ_k代表了启发函数，其值可由式$\delta_k=C_{\mathrm{re}}(k)-C_{\mathrm{pr}}(f_{ij})$可得，其中，$C_{\mathrm{re}}(k)$表示当前$k$节点剩余的资源，$C_{\mathrm{pr}}(f_{ij})$表示$f_{ij}$对资源的最低需求预测结果。信息素的更新为

$$\zeta_{\mathrm{ck}}(\mathrm{iN}+1)=(1-\rho\%)\zeta_{\mathrm{ck}}(\mathrm{iN})+\Delta\zeta_{\mathrm{ck}}(\mathrm{iN}) \tag{19}$$

$$\Delta\zeta_{\mathrm{ck}}(\mathrm{iN})=\sum_{s=1}^{\mathrm{aN}}\Delta\zeta_{\mathrm{ck}}^{s}(\mathrm{iN}) \tag{20}$$

$$\Delta\zeta_{\mathrm{ck}}^{s}(\mathrm{iN})=\begin{cases}C_{\mathrm{ac}}(s)/Q, & \text{第}s\text{只蚂蚁在本次路径选择中依次经过了节点}c,k\\ 0, & \text{其它}\end{cases} \tag{21}$$

ρ为信息素挥发因子，$C_{\mathrm{ac}}(s)$表示第s只蚂蚁路径选择完成之后计算出的服务功能链所分配的计算资源总和；Q为常系数，iN为迭代次数，aN表示蚂蚁规模。为了防止局部最优的出现，本文将各条寻优路径上的信息素浓度限制在$[\zeta_{\min},\zeta_{\max}]$，即按照式(22)对信息素浓度进行约束

$$\zeta_{\mathrm{ck}}(\mathrm{iN}+1)=\begin{cases}\zeta_{\max}, & \zeta_{\mathrm{ck}}(\mathrm{iN})>\zeta_{\max}\\ \zeta_{\mathrm{ck}}(\mathrm{iN}), & \zeta_{\min}<\zeta_{\mathrm{ck}}(\mathrm{iN})<\zeta_{\max}\\ \zeta_{\min}, & \zeta_{\mathrm{ck}}(\mathrm{iN})<\zeta_{\min}\end{cases} \tag{22}$$

4 仿真与性能分析

本文将VNFs的平均端到端时延、平均比特丢失率等作为评价指标，同时VNF的调度及资源分配上，为了更好地体现MMACA算法的性能，对比了基于普通的蚁群算法静动态的资源分配方案(ACA-D、ACA-S)，以及文献[5]中基于遗传算法的VNF部署及资源动态分配方案(GA-D)。在资源需求的预测问题上，对比了文献[15]基于ARMA的负载预测算法。本文仿真的过程中假设的网络场景为全连接型网络，网络规模包含8个VM和3种网络切片业务。同时假设实现每一种网络切片业务的服务功能链由固定且相等的VNF个数组成，并且每一VNF至少能在3～5个VM上执行。相关的仿真参数设置如表4所示。

表 4 仿真参数

仿真参数	仿真值
网络切片业务数量	3
短周期时长	10 s
长周期时长	4 h
数据包到达过程	泊松分布
数据包到达速率	[50, 100]个/s
数据包大小	指数分布
队列缓存大小	10 MB
数据训练窗口	8
神经元个数	10
仿真时间	24 h
学习率	0.01

本文每个长周期内用户请求切片业务数据包的到达速率λ_i服从均值为[50，100]个/s的泊松分布，每个数据包的大小服从均值为500 kbit的指数分布，同时假设每个数据包被VNF模块处理后大小不发生变化。为了模拟业务到达以天为单位的周期性循环，业务到达模式以每经过140000个请求循环[9]。

接下来首先对本文所提方案中切片未来负载的预测能力进行验证。为了验证模型的预测效果，本文采用平均绝对百分比误差(Mean Absolute Percentage Error, MAPE)对预测的准确度进行衡量，MAPE定义为

$$\mathrm{MAPE}=\left(\frac{1}{\Omega}\sum_{t=1}^{\Omega}\left|\frac{o_i(t)-\xi_i(t)}{\xi_i(t)}\right|\right) \tag{23}$$

其中，$o_i(t)$为预测值，$\xi_i(t)$为真实值，Ω为测试样本的数量。MAPE的值越小，说明预测的准确度越高。图3和图4刻画了在6个连续长周期下数据包处理后切片队列状态的真实情况与基于不同预测方案所得值的拟合效果。图5表示不同方案对最大负载预测误差对比，可以看出基于LSTM的预测方案整体效果优于ARMA。图6表示LSTM训练过程的平

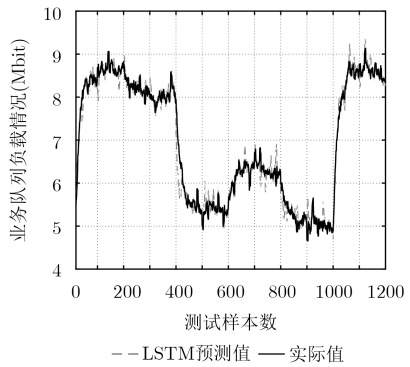

图 3 切片业务队列负载情况LSTM预测值与实际值对比图

均误差，从图6中可知初始训练阶段总平均误差处于较高水平，经过约300次迭代之后，误差基本平稳并控制在较低范围。

图7显示了不同长周期内不同方案下的平均端到端时延对比图，可以看出本文提出的MMACA算法在性能上与其它3种方案相比处于较高水平，这是因为MMACA算法采用了基于流量感知的资源分配机制，提前进行针对性的VNF调度并按需以比例公平的方式进行资源配置，并在SFC部署过程中通过参数动态自适应的改进策略，降低了算法陷入局部最优解的可能性，提高了SFC部署的寻优能力。而其它3种资源配置方案存在算法性能上的不足，或忽略了网络业务动态变化这一特点，在某种程度上会造成资源配置与实际需要的不匹配问题。图8显示了在6个不同的长周期下，不同方案在比特丢失率上的性能评估。可以看出，当长周期的数据包到达率处于较低水平时，所以各方案的比特丢失率都处于较为理想的水平。当长周期的数据包到达率增大时，各VNF开始竞争有限的CPU资源，此时由于MMACA方案合理地按需调度有限的资源，所以采用该算法的VNF调度及资源分配方案在平均比特丢失率上依然保持较高的性能，而由于GA-D和ACA-D方案容易陷入局部最优解，所以出现了相对于MMACA方案较为严重的比特丢失情况。而ACA-S方案采用的是静态的资源调度方案，在比特丢失率上的性能表现得十分不理想。

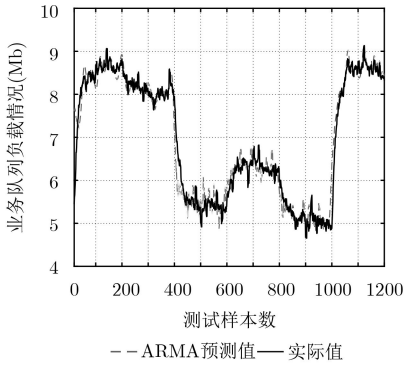

图 4 切片业务队列负载情况ARMA预测值与实际值对比图

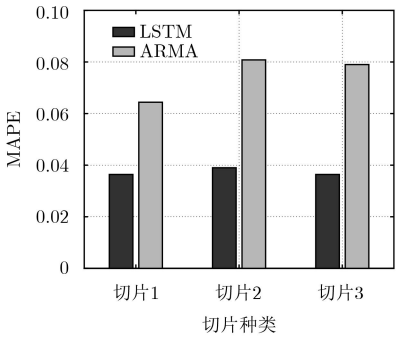

图 5 预测误差对比图

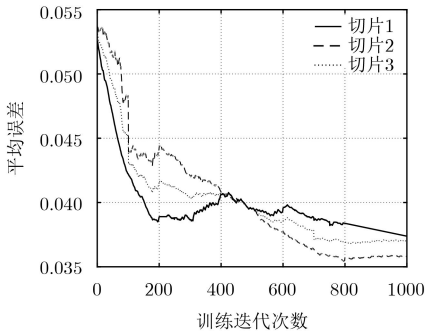

图 6 LSTM训练过程平均误差

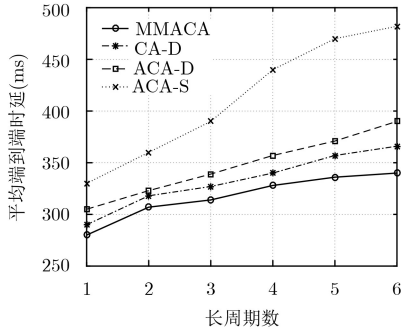

图 7 平均端到端时延的对比图

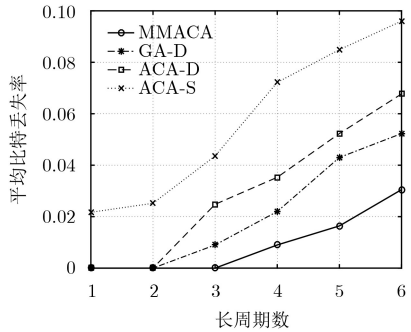

图 8 不同方案平均比特丢失率对比图

5　结束语

针对无线虚拟化网络中在时间域上用户请求业务的变化性和信息反馈的时滞效应而引起虚拟资源分配的不合理，本文提出了一种基于预测的虚拟网络功能部署算法MMACA，该算法将计算资源作为虚拟化的载体，考虑到变化的业务请求引起数据的积累而产生的队列积压，提出了一种基于LSTM神经网络的流量感知模型，并利用SFC最低资源需求的预测结果设计了一种基于自适应蚁群算法的虚拟网络功能调度方案。仿真结果表明，本文的预测算法精确度很高，同时，将预测与虚拟网络功能结合的方案可有效地降低网络切片业务的整体端到端时延和比特丢失率。

参 考 文 献

[1] MAHMOOD N H, LAURIDSEN M, BERARDINELLI G, *et al*. Radio resource management techniques for eMBB and mMTC services in 5G dense small cell scenarios[C]. IEEE 84th Vehicular Technology Conference, Montreal, Canada, 2016: 1–5. doi: 10.1109/VTCFall.2016.7881187.

[2] VASSILARAS S, GKATZIKIS L, LIAKOPOULOS N, *et al*. The algorithmic aspects of network slicing[J]. *IEEE Communications Magazine*, 2017, 55(8): 112–119. doi: 10.1109/MCOM.2017.1600939.

[3] HERRERA J G and BOTERO J F. Resource allocation in NFV: A comprehensive survey[J]. *IEEE Transactions on Network & Service Management*, 2016, 13(3): 518–532. doi: 10.1109/TNSM.2016.2598420.

[4] SALLENT O, PEREZ-ROMERO J, FERRUS R, *et al*. On radio access network slicing from a radio resource management perspective[J]. *IEEE Wireless Communications*, 2017, 24(5): 166–174. doi: 10.1109/MWC.2017.1600220WC.

[5] LONG Q, ASSI C, and SHABAN K. Delay-aware scheduling and resource optimization with network function virtualization[J]. *IEEE Transactions on Communications*, 2016, 64(9): 3746–3758. doi: 10.1109/TCOMM.2016.2580150.

[6] HUANG Huang, GUO Song, WU Jinsong, *et al*. Service chaining for hybrid network function[J]. *IEEE Transactions on Cloud Computing*, 2017. doi: 10.1109/TCC.2017.2721401.

[7] ZHU Qixuan and ZHANG Xi. Game-theory based buffer-space and transmission-rate allocations for optimal energy-efficiency over wireless virtual networks[C]. IEEE Global Communications Conference, San Diego, USA, 2015: 1–6. doi: 10.1109/GLOCOM.2015.7417845.

[8] FENG Qiu, ZHANG Bin, and GUO Jun. A deep learning approach for VM workload prediction in the cloud[C]. 201617th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Shanghai, China, 2016: 319–324. doi: 10.1109/SNPD.2016.7515919.

[9] MIJUMBI R, HASIJA S, DAVY S, *et al*. Topology-aware prediction of virtual network function resource requirements[J]. *IEEE Transactions on Network & Service Management*, 2017, 14(1): 106–120. doi: 10.1109/TNSM.2017.2666781.

[10] AGARWAL S, MALANDRINO F, CHIASSERINI C, *et al*. Joint VNF placement and CPU allocation in 5G[C]. IEEE INFOCOM 2018 - IEEE Conference on Computer Communications, Honolulu, USA, 2018: 1943–1951. doi: 10.1109/INFOCOM.2018.8485943.

[11] ZHANG Haijun, LIU Na, CHU Xiaoli, *et al*. Network slicing based 5G and future mobile networks: mobility, resource management, and challenges[J]. *IEEE Communications Magazine*, 2017, 55(8): 138–145. doi: 10.1109/MCOM.2017.1600940.

[12] YANG Jian, ZHANG Shuben, WU Xiaomin, *et al*. Online learning-based server provisioning for electricity cost reduction in data center[J]. *IEEE Transactions on Control Systems Technology*, 2017, 25(3): 1044–1051. doi: 10.1109/TCST.2016.2575801.

[13] CHENG Aolin, LI Jian, YU Yuling, *et al*. Delay-sensitive user scheduling and power control in heterogeneous networks[J]. *IET Networks*, 2015, 4(3): 175–184. doi: 10.1049/iet-net.2014.0026.

[14] GLOROT X and BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[J]. *Journal of Machine Learning Research*, 2010, 9: 249–256.

[15] 唐伦, 杨希希, 施颖洁, 等. 无线虚拟网络中基于自回归滑动平均预测的在线自适应虚拟资源分配算法[J]. 电子与信息学报, 2019, 41(1): 16–23. doi: 10.11999/JEIT180048.
TANG Lun, YANG Xixi, SHI Yingjie, *et al*. ARMA-prediction based online adaptive dynamic resource allocation in wireless virtualized networks[J]. *Journal of Electronics & Information Technology*, 2019, 41(1): 16–23. doi: 10.11999/JEIT180048.

唐 伦：男，1973年生，教授，博士生导师，主要研究方向为新一代无线通信网络、异构蜂窝网络、软件定义无线网络等.

周 钰：男，1993年生，硕士生，研究方向为5G网络切片资源分配和深度学习.

杨友超：男，1993年生，硕士生，研究方向为网络虚拟化和切片资源分配.

赵国繁：女，1993年生，硕士生，研究方向为5G网络切片中的资源分配，可靠性.

陈前斌：男，1967年生，教授，博士生导师，主要研究方向为个人通信、多媒体信息处理与传输、下一代移动通信网络.