虚拟网络功能资源容量自适应调整方法

袁 泉* 游 伟 季新生 汤红波
(解放军战略支援部队信息工程大学 郑州 450002)

摘 要：为了实现网络功能虚拟化平台中物理资源的动态按需分配，该文提出一种虚拟网络功能资源容量自适应调整方法。该方法首先利用长短期记忆网络预测平台流量的变化趋势，然后结合流量预测结果设计了一种基于多层前馈神经网络的虚拟网络功能资源需求预测方法，最后根据资源需求预测结果，设计了一种基于动态编码遗传算法的虚拟网络功能动态部署方法，实现虚拟网络功能资源容量的自适应调整。实验结果表明，与现有的资源容量调整方法相比，该文提出的资源容量自适应调整方法能够降低流量预测误差对资源需求预测结果的影响，降低资源需求预测的相对误差，减少虚拟网络功能实例占用的服务器数量。

关键词：网络功能虚拟化；虚拟网络功能资源分配；前馈神经网络；整数线性规划；遗传算法
中图分类号：TN915; TP393 **文献标识码**：A **文章编号**：1009-5896(2021)07-1841-08
DOI: 10.11999/JET200110

Adaptive Scaling of Virtualized Network Function Resource Capacity

YUAN Quan YOU Wei JI Xinsheng TANG Hongbo
(*PLA Strategic Support Force Information Engineering University*, *Zhengzhou* 450002, *China*)

Abstract: In order to realize on-demand physical resource allocation in network function virtualization platform, an adaptive virtualized network function scaling method is proposed. The proposed method first use long short term memory network to realize traffic forecasting. Then combining with the forecasting result, a forward neural network-based approach is designed to predict resource demand of requested virtualized network function. Finally, according to the result of resource demand prediction, a dynamic encoding genetic algorithm is proposed to realize dynamic deployment of virtualized network function instances. The experiment results show that compared with existing scaling methods, the proposed scaling method can reduce the negative impact of inaccurate traffic forecasting, decrease the relative error of resource demand prediction as well as the total number of servers occupied by requested virtualized network function instances.

Key words: Network Function Virtualization (NFV); Virtualized Network Function (VNF) resource allocation; Feedforward neural network; Integer linear programming; Genetic algorithm

1 引言

网络功能虚拟化(Network Function Virtualization, NFV)实现了传统网络功能与专用硬件的解耦，从专用网络设备中抽象出软件化的虚拟网络功能(Virtualized Network Function, VNF)，推动了网络服务部署模式的根本性转变——由僵化的“网元”部署模式转向相对灵活的切片部署模式[1]。在基于网络切片的部署模式中，服务提供商通过创建一组有序的VNF集合为租户提供定制化的网络服务，并通过NFV管理和编排模块(Management And Orchestration Module, MANO)动态调整VNF集合占用的资源容量，实现物理资源(计算、内存和磁盘资源等)的动态按需分配[2]。以物理资源“按需分配”为目标，文献[3]提出了VNF资源容量调整的概念，并设计了基于VNF重映射的解决方案。文献[4]总结了现有的VNF资源容量调整方法，并将其分为以下两种类型[4]：(1)基于流量预测的调整方法[5–10]，通过预测VNF集合输入流量的变化趋势，预判其中VNF实例数量的变化，提前部署或移除相应的VNF实例实现资源容量调整。(2)基于在线优化理论的调整方法[11–14]，在流量不可预测的场景中，利用在线优化模型(如Ski-rental模型)求解局部最优的容量调整策略。本文主要针对基于流量

收稿日期：2020-02-17；改回日期：2020-10-05；网络出版：2020-12-14
*通信作者： 袁泉 b101180153@smail.nju.edu.cn
基金项目：国家自然科学基金(61801515)，国家自然科学基金创新群体项目(61521003)
Foundation Items: The National Natural Science Foundation of China (61801515), The National Natural Science Foundation Innovative Groups Project of China (61521003)

预测的VNF资源容量调整方法进行研究。

文献[5]分析并提取了运营商云数据中心网络的流量特征，设计了一种基于深度神经网络的流量预测算法，并结合预测数据提出了一种基于整数规划的VNF资源容量调整方法。文献[6]提出了一种面向5G核心网上行链路场景的流量预测方法，该方法通过分析流入无线接入和移动性管理功能(Access and Mobility management Function, AMF)上行链路的流量数据提取样本集，然后设计和训练循环神经网络实现流量预测。文献[7]提出了一种轻量级的流量预测方法，引入K均值和蒙特卡罗方法加快循环神经网络的训练过程，提高训练效率，降低预测方法的时间复杂度。文献[8]针对流量预测准确度较低的场景提出了一种预测误差补偿方法，该方法利用在线学习算法最小化流量预测误差对VNF资源容量调整策略的影响，降低了由预测误差带来的容量调整开销。文献[9]提出了一种基于排队模型的VNF资源调整方法，利用长短期记忆网络(Long Short Term Memory, LSTM)预测流量变化趋势，并设计了一种基于最大最小蚁群算法的VNF动态部署方法，实现计算资源利用率最大化。文献[10]提出了VNF资源容量调整过程中VNF性能和资源占用之间的均衡问题，并设计了基于排队模型的启发式算法搜索最优解，在保证VNF性能的基础上最小化VNF集合占用的物理资源。综合分析上述基于流量预测的VNF资源容量调整方法，其中电信网和云数据中心场景下的流量预测算法已经有了相对完备的研究成果，但是相应的VNF部署方法仍存在两方面缺陷：(1)无法根据流量预测结果准确预测VNF的资源需求。现有模型主要基于相关函数法[5–7,11]建立流量预测结果与VNF资源需求之间的映射关系，但是由于许多类型的资源利用存在长相关、重尾、非线性和多尺度等特征，上述方法难以根据流量预测结果准确预测VNF的资源需求。(2)在容量调整的过程中未考虑如何减少VNF实例占用的服务器数量。现有研究主要针对大规模电信云数据中心场景下的VNF容量调整问题，其系统模型中采用cross rack pipelined SFC模式[5]部署服务功能链，即在某一机框中仅部署单一类型的VNF实例，不同VNF之间通过机框外部的物理链路相互串联，组成服务功能链为租户提供服务。cross rack pipelined SFC模式的优点在于可以为VNF的扩容提供充足物理资源，且便于VNFM统一管理编排各个类型的VNF。但是对于中小规模的数据中心，由于服务器数量较少，数据中心资源规划相对紧张，云平台无法为各个类型的VNF分配独立的机框资源，服务提供商通常选择SFC run-to-complete in a rack模式部署SFC，即在同一机架内完成一条SFC的部署。由于该模式下的服务器资源相对短缺，需要研究如何在SFC run-to-complete in a rack模式中协同部署不同规格的VNFC实例，减少SFC占用的服务器数量，提高云平台的资源可用性。

针对已有研究存在的问题，本文提出一种VNF资源容量自适应调整方法，并采用trace-dirven方法[15]进行了仿真实验。该方法将真实云环境中获取的一系列观测数据输入仿真程序，保证了实验结果的有效性。首先，我们从实验云数据中心的接入网关中获取输入流量数据，并利用LSTM网络预测其变化趋势。然后在OpenStack日志文件中统计相应时段各规格VNF实例的历史部署数据，结合流量预测结果设计多层前馈神经网络预测租户的VNF资源需求。该方法综合考虑了流量变化、资源需求变化和业务类型等因素对资源预测准确度的影响，能够降低VNF资源需求预测的相对误差。最后基于VNF资源需求预测数据，建立改进的二次分配模型描述VNF动态部署问题，并设计了一种动态编码遗传算法求解该NP难问题。该算法能够根据资源需求的变化自适应地调整VNF的部署策略，减少VNF实例部署占用的服务器数量，提高物理资源的可用性。

2 系统模型与优化目标

2.1 服务链部署模型和VNF资源容量调整问题

服务功能链是一组有序的VNF集合，在NFV平台上，服务提供商利用服务功能链技术向租户提供端到端定制化服务。如图1所示，服务功能链的基本组件包含分类器、转发器和VNF集合3部分。其中，分类器主要负责业务流量的接入控制和路由策略管理，转发器负责流量解析、封装和转发，VNF集合负责业务流量处理[16]。租户请求的业务流量通过分类器进入服务功能链，根据分类器中的路由策略被转发至相应的转发器，再由转发器路由至VNF集合进行业务数据处理，完成数据处理后业务数据将被依次转发至目的端分类器，最终接入目的服务器。VNF资源容量调整问题的研究目标：根据流经服务功能链的实时流量自适应地调整VNF集合内实例的规格和数量，实现物理资源的按需分配，提高资源利用率。如图1所示，VNF资源容量调整问题可分为两个阶段：(1)生成资源需求，根据流经服务功能链的流量数据预测未来的VNF资源需求，生成VNF资源需求视图，明确下一个时间段需要生成的各个规格VNF实例的数量。(2)VNF部

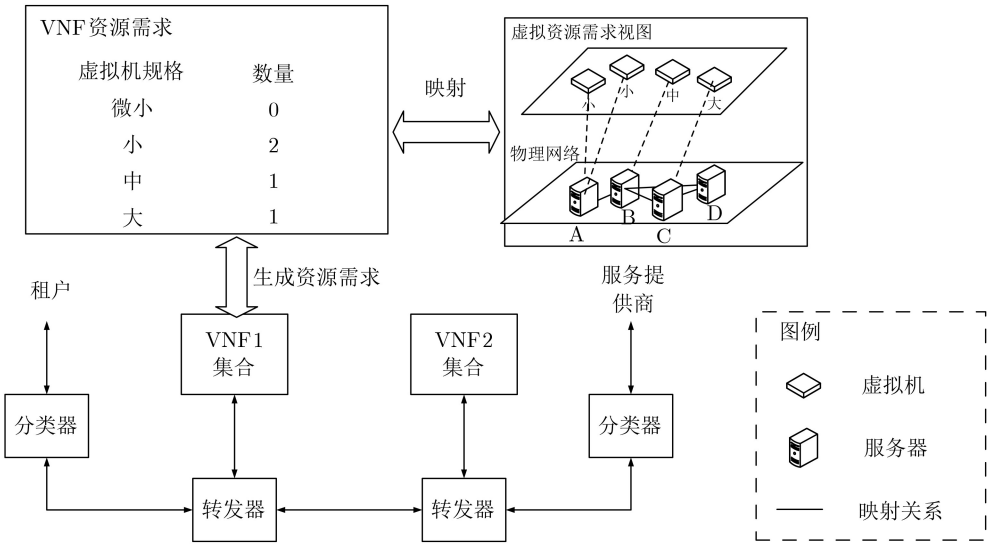

图 1 服务功能链模型和VNF资源容量调整

署(映射)，根据当前时刻物理网络的全局视图，将VNF虚拟资源需求视图映射到底层物理网络，完成VNF实例化过程。

2.2 优化目标

为了提高NFV平台物理资源的利用效率，本文针对VNF资源调整的两个阶段，提出了相应的优化目标：(1)在生成资源需求阶段，降低流量预测误差对资源需求预测结果的影响，提高VNF资源需求预测的准确度。(2)在动态部署阶段，实现VNF集合中所有实例的集中化部署，减少VNF占用的服务器数量。

3 VNF资源需求预测方法

为了提高VNF资源需求预测的精度，本节首先利用LSTM网络进行流量预测，然后提出了一种基于多层前馈神经网络的资源需求预测方法，建立流量数据和资源需求数据之间的映射关系模型，实现资源需求精确预测。

3.1 基于LSTM网络的流量预测

根据文献[5]在运营商云数据中心获取的测试数据，以OpenStack为代表的虚拟化云平台完成虚拟机创建和迁移需要“分钟级”的准备时间，而5G网络服务的QoS要求运营商将端到端时延降低到毫秒级[17]。在此背景下，依靠网络监控功能(如OpenStack Ceilmeter)上报的流量数据，实时调整VNF资源容量的“反应式”部署方法无法满足相关业务的时延需求。因此，运营商亟需提高网络流量预测的准确度，以便根据流量预测数据提前完成VNF部署。

本节采用了目前比较成熟的时间序列预测方法LSTM网络模型进行流量预测。LSTM网络是基于长短期记忆的循环神经网络，引入了长短期记忆细胞结构代替一般循环神经网络中普通的隐层神经元，可以动态改变当前时间步细胞结构中输入、状态和输出信息对应的权重，在时间尺度上动态调节网络中历史输入数据对预测结果的影响，解决循环神经网络中的长期依赖问题。图2为LSTM网络中的长短期记忆“细胞”框架[18]，其主要功能通过3个门结构实现：(1)遗忘门，通过动态改变当前时间步状态信息对应的权值决定如何从当前细胞状态中丢弃历史状态信息，实现短期记忆功能。(2)输入门，通过改变前序时间步状态信息对应的权值决定如何向当前细胞中添加历史状态，实现长期记忆功能。(3)输出门，通过改变当前时间步输出信息对应的权值，决定如何输出当前时间步预测信息，实现时间序列数据预测。

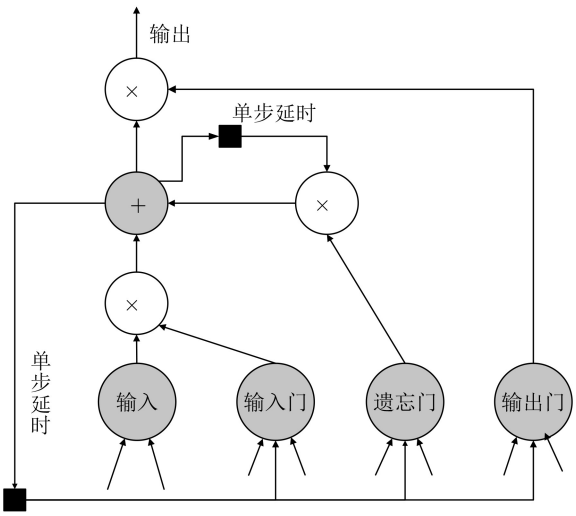

图 2 LSTM循环网络“细胞”框架

3.2 基于多层前馈神经网络的资源需求预测

如图3所示，前馈神经网络(Feedforward Neural Network, FNN)采用单向多层的网络结构，整个网络可以用一个有向无环图表示[18]。网络中每一层包含多个神经元且各层神经元仅能接受前一层神经元传递的信号，其中第0层为输入层，包含的神经元个数与输入数据特征的维度相等，最后一层为输出层，包含的神经元个数与输出向量的维度相等，其他为隐层。文献[19]证明，只需1个包含足够多神经元的隐层，FNN就能以任意精度逼近任意复杂度的连续函数。借助其强大的回归预测能力，本节通过设计和训练多层FNN建立流量数据与资源需求数据之间的映射模型。

本文所用数据集采用trace-driven方法从实验室云数据中心获取，该云平台包含7台华为RH2288H V3型高性能服务器，1台传统交换机和3台盛科V580-32X型SDN交换机。表1列出了数据集中的输入特征。其中$\boldsymbol{L}(t)$为接入网关服务器的历史流量向量，其中任意元素$l(t-i)$表示$t-i$时刻的历史流量。$p(t+1)$表示$t+1$时刻在该网关处的预测流量，ServerID表示当前接入网关服务器的序号，$\boldsymbol{R}(t)$表示t时刻的资源需求向量，其中任意元素$r_j(t)$表示t时刻平台实例化的第j种规格VNF的数量。设平台可实例化m种规格的VNF实例，则资源需求线路可表示为$\boldsymbol{R}(t)=\{r_1(t),\cdots,r_j(t),\cdots,r_m(t)\}$。FNN输出层信号为$t$+1时刻VNF资源需求向量$\widehat{\boldsymbol{R}}(t+1)=\{\widehat{r}_1(t+1),\widehat{r}_2(t+1),\cdots,\widehat{r}_j(t+1)\}$，其中任意元素$\widehat{r}_j(t+1)$表示$t$+1时刻所需第$j$种规格VNF数量的预测值。

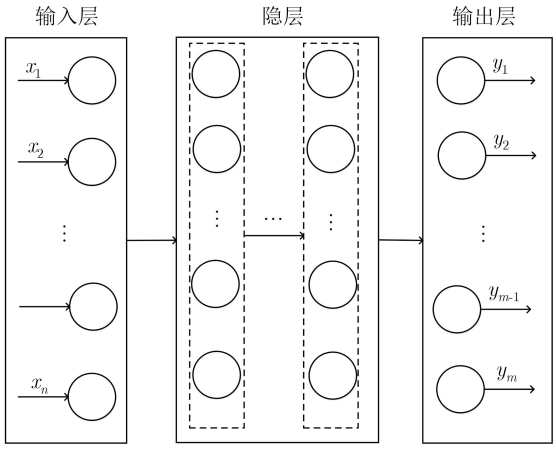

图 3 多层前馈神经网络结构

4 VNF动态部署

根据多层FNN预测的VNF资源需求，本节提出了一种基于遗传算法的VNF部署算法，旨在实现VNF需求视图中所有实例的集中化部署，最小化VNF集合占用的物理服务器数量，以便运营商将整条服务功能链部署到同一机框中管理(SFC run-to-complete in a rack模式[5])。

在VNF部署模型中，数据中心网络可表示为无向赋权图$G_S=(N_S,L_S)$，其中N_S表示高性能服务器集合，L_S表示服务器间链路组成的集合。VNF请求可表示为无向赋权图$G_V=(N_V,L_V)$，其中N_V表示VNF实例组成的集合，L_V表示VNF实例间链路组成的集合，部署过程可抽象为映射$f:G_V\to G_S$。设服务器的数量为n，服务器提供的资源类型共有k类(CPU、内存、硬盘等)，云平台可实例化m种规格的VNF。定义$n\times k$维矩阵$\boldsymbol{C}(t)$为t时刻服务器资源容量矩阵，其中任意元素$c_{ij}(t)$表示t时刻服务器i可提供的第j类资源的容量。m维列向量$\boldsymbol{R}(t)$为t时刻VNF资源需求向量，其中任意元素$r_j(t)$表示t时刻平台需要实例化的第j种规格VNF的数量。$m\times k$维矩阵$\boldsymbol{V}$表示实例化不同规格的VNF所需的物理资源，其第i行所有元素组成的行向量表示实例化第i种规格的VNF所需的各类型物理资源的数量，元素v_{ij}表示实例化第i种规格的VNF所需第j类物理资源的数量。$n\times n$维矩阵$\boldsymbol{E}(t)$为t时刻服务器邻接矩阵，其中元素e_{ij}表示服务器i和服务器j之间的有效带宽。$m\times m$维矩阵$\boldsymbol{D}$表示VNF邻接矩阵，其中元素d_{ij}表示i规格的VNF到j规格的VNF通信链路的带宽开销。$n\times m$维矩阵$\boldsymbol{X}(t)$表示t时刻VNF部署决策，其中元素$x_{ij}=\alpha$表示有α个j规格的VNF部署在物理服务器i上。根据上述模型，VNF动态部署问题可表示为式(1)—式(5)描述的整数线性规划问题。

优化目标

表 1 t时刻资源需求预测特征提取

特征参数	意义
$\boldsymbol{L}(t)$	历史流量向量，包含$t-i$时刻到t时刻的实测流量数据$\boldsymbol{L}(t)=\{l(t-i),\cdots,l(t)\}$
$p(t+1)$	$t+1$时刻的流量预测数据
ServerID	当前流量流经的接入网关服务器序号，用于区分
$\boldsymbol{R}(t)$	历史资源需求向量，包含t时刻的资源需求数据$\boldsymbol{R}(t)=\{r_1(t),\cdots,r_j(t),\cdots,r_m(t)\}$，其中任意元素$r_j(t)$表示$t$时刻平台实例化第$j$种规格VNF的数量

$$\min \sum_{t} \sum_{i=1}^{n} \varepsilon \left(\sum_{j=1}^{m} x_{ij}(t) - 1 \right) \tag{1}$$

约束条件

$$\boldsymbol{X}(t) \cdot \boldsymbol{V} \le \boldsymbol{C}(t) \tag{2}$$

$$\boldsymbol{X}^{\mathrm{T}}(t) \cdot \boldsymbol{I}_{n\times 1} = \boldsymbol{R}(t) \tag{3}$$

$$\sum_{j=1}^{m} \sum_{q=1}^{m} x_{ij}(t) \cdot x_{pq}(t) \cdot d_{jq} \le e_{ip}(t), \forall i, p \in N_S, i \ne p \tag{4}$$

$$\forall x_{ij}(t) \in N \tag{5}$$

式(1)为VNF动态部署的优化目标，其中$\varepsilon(\cdot)$表示阶跃函数，若某一服务器上部署了VNF实例则$\varepsilon\left(\sum_{j=1}^{m} x_{ij}(t) - 1\right) = 1$，反之$\varepsilon\left(\sum_{j=1}^{m} x_{ij}(t) - 1\right) = 0$。该式提出的优化目标表示在所有时刻部署相应VNF资源需求所需的最小服务器数量。式(2)为服务器资源容量约束，表示部署在任一服务器上的所有VNF实例所占用的各类型物理资源不能超过当前时刻该服务器能提供的资源容量。式(3)为VNF资源需求约束，其中$\boldsymbol{I}_{n\times 1}$为元素值全部为1的$n$维列向量，该式表示对于任意规格VNF，已部署的实例数量之和等于当前时刻资源需求视图中该规格实例的数量。式(4)为链路带宽约束，表示映射到任一物理链路的VNF链路带宽之和不能超过当前时刻该物理链路的带宽容量。式(5)定义了决策变量的定义域，其中N为自然数集合。综上，式(1)—式(5)定义的VNF动态部署问题可规约为扩展的二次分配问题(Quadratic Assignment Problem, QAP)，文献[20]已证明该问题为NP难问题。

为了降低求解的计算时间复杂度，本节设计了一种基于动态编码遗传算法的VNF动态部署算法(Dynamic-encoding Genetic Algorithm, DG Alg.)，具体流程如表2所示。步骤(2)和步骤(3)能够根据当前时刻请求中VNF实例的数量动态编码染色体，式(6)表示编码后的染色体，其中第1到r_1个元素的值对应部署第1类规格VNF实例的服务器序号，例如：$n_1 = 2$表示第1类规格的第1个VNF实例部署在了2号服务器上，后续编码以此类推。步骤(4)—步骤(11)计算了种群中各染色体适应度，其中步骤(5)将种群中任一染色体i上的基因分割为长度为$r_1, r_2, \cdots, r_m$的m个子序列，每一个子序列包含了一种规格的VNF实例与物理服务器之间的映射关系，通过统计各个子序列中服务器的序号获取当前染色体对应的部署决策矩阵$\boldsymbol{X}_i(t)$。步骤(7)通过式(1)计算每个染色体中VNF实例部署占用的服务器数量，并将其倒数作为染色体对应的适应度。进化过程中算法将保留适应度高的染色体，选出种群中占用服务器数量最少的部署方案。步骤(12)—步骤(17)为进化和自然选择过程，其中步骤(13)实现了染色体排序和基因交叉遗传，首先将种群中所有染色体按照适应度排序，再归一化适应度序列获得各染色体的候选概率。根据候选概率选择种群中的染色体，对其每一个基因按照概率P_x进行交叉操作，交叉后获得子代染色体。步骤(14)为变异操作，按照概率P_m将染色体上的基因n_i 替换为新的基因n_i'，变异操作可以在种群中引入新的基因，防止算法陷入局部最优。

$$[n_1, \cdots, n_{r_1}, n_{r_1+1}, \cdots, n_{r_1+r_2}, \cdots, n_{N_g(t)}] \tag{6}$$

5　性能评估及分析

5.1　实验环境和参数设置

(1) 实验环境和数据集采集：实验云数据中心基于OpenStack搭建，包含7台华为RH2288H V3型服务器，其中2台高配服务器，5台低配服务器，具

表 2　VNF动态部署算法(DG Alg.)

输入： t时刻资源容量需求向量$\boldsymbol{R}(t)$；服务器资源容量矩阵 $\boldsymbol{C}(t)$；服务器邻接矩阵$\boldsymbol{E}(t)$

输出： 最优VNF部署决策矩阵$\boldsymbol{X}^*(t)$

(1) 初始化遗传算法参数：种群规模NIND，最大遗传代数MAXGEN，交叉和变异概率P_x, P_m

(2) 计算t时刻染色体上的基因数目 $N_g(t) = \sum_{i=1}^{m} r_i(t)$

(3) 随机初始化种群中的每个染色体，每个基因位的进制设为服务器数量n

(4) for $i = 1$: NIND

(5) 根据$\boldsymbol{R}(t)$分段统计各基因位的服务器序号，计算染色体 n对应的VNF部署决策变量$X_i(t)$

(6) if $X_i(t)$满足约束条件式(2)-式(5)do

(7) 根据式(1)计算当前染色体的适应度

(8) else

(9) 当前染色体对应的部署策略无法完成部署，适应度为惩罚值N_p

(10) end if

(11) end for

(12) for gen =1 : MAXGEN

(13) 根据适应度计算染色体参与遗传的优先级，以概率P_x对候选染色体上基因进行交叉遗传

(14) 按照概率P_m选择染色体上任意基因进行变异操作

(15) 获取子代种群，重复步骤(4)—步骤(11)，计算子代种群各个染色体的适应度

(16) 用子代中适应度高的染色体替换父代中适应度低的染色体，形成新的种群

(17) end for

(18) 保留最终代中最优染色体，返回其对应的VNF部署决策变量$\boldsymbol{X}^*(t)$

体配置参数如表3所示。网络设备包含1台10 GB传统交换机和3台盛科V580-32X型SDN交换机。本文实验使用的流量数据集采用trace-driven方法从数据中心接入网关服务器获取，服务器datastore UUID为5c2f4446-08a8fa88-6631-289e97db65c6，数据集中流量数据统计的观测时间间隔为1 h，观测时长为1230 h，图4展示了观测时长内各个时刻的流量大小。承载VNF实例的虚拟机规格在OpenStack平台上预置为4类，具体参数如表4所示，VNF资源需求数据集通过统计OpenStack虚拟机创建日志获取。

(2) 实验参数设置：采用PyTorch框架搭建LSTM网络和多层FNN，经调试后试验参数选取如下：(1)基于LSTM的流量数据预测：时间步长设为24，学习率为0.01，学习周期为10000，损失函数为最小均方误差，优化器选择Adam，选取70%的流量数据作为训练集，30%的数据用作测试集。(2)资源需求数据预测：为了验证本文资源需求预测方法有效性，实验中将本文设计的多层FNN方法与基于LSTM的资源需求预测方法进行了对比。实验选取70%的资源需求数据作为训练集，30%的数据作为测试集。多层FNN输入特征中的历史流量向量$\boldsymbol{L}(t)$选取过去6个时间步的历史流量数据，网络包含3个隐层其结构为24-36-24，批大小设为120，批学习周期为20，学习率为0.05，损失函数为最小均方误差，优化器选择SGD。对比试验中，LSTM网络时间步长设为24，学习率为0.01，学习周期为20000，损失函数为最小均方误差，优化器选择Adam。(3)动态编码遗传算法：种群大小为100，交叉概率为0.7，变异概率为0.3，最大进化代数为100。

表 3 服务器参数

型号	CPU	内存(GB)	硬盘(TB)
高配	28 CPUs x Intel(R) Xeon(R) CPU E5-2660 v4 @ 2.00GHz	128	14
低配	20 CPUs x Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz	64	14

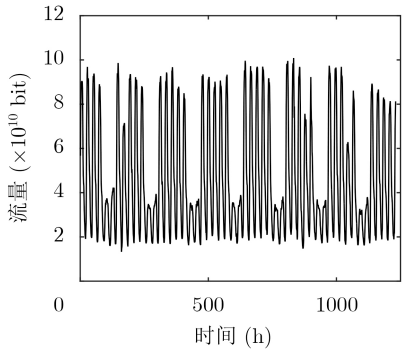

图 4 数据集流量

表 4 虚拟机规格参数

规格	VCPU(个)	内存(GB)	硬盘(GB)
微小	1	1	20
小	2	2	50
中	4	4	200
大	4	8	500

5.2 结果分析

图5为各时刻LSTM网络流量预测的相对误差，预测流量序列与原始流量序列之间的均方误差为1.0751×10^{18}。预测误差结果表明，针对该样本集，LSTM网络虽然能够预测流量变化的基本趋势，但是在全时间尺度上仍会出现预测误差较大的情况。为了验证本文所提多层FNN资源需求预测方法的有效性，本文将其与现有的LSTM资源需求预测方法进行了对比。图6至图9给出了两种方法在1230 h内各个时刻的相对误差散点图，其中图6至图8中个别时刻相对误差较大的原因是该时刻VNF需求视图中仅请求了1个该规格的VNF实例。对比LSTM方法与多层FNN方法资源需求预测的相对误差可以得出结论，本文提出的多层FNN预测方法可以有效降低LSTM网络流量预测误差对VNF资源需求预测的影响，降低各个规格VNF实例数量预

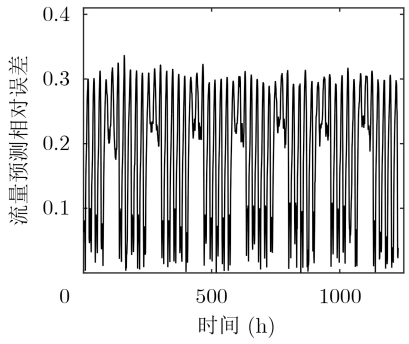

图 5 流量预测结果

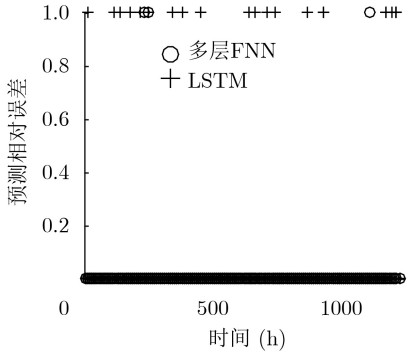

图 6 微小规格VNF实例预测误差

测的相对误差，提高VNF实例资源需求预测的精确度。

图10对比了3种方法在全时间尺度上的累计服务器占用数量，其中Cplex是IBM开发的整数规划求解工具，在网络规模较小的场景中能够获得本文提出模型的理论最优解，Greedy算法是目前OpenStack内置的集中化部署算法。相比于Greedy算法，本文提出的DG Alg.能够明显降低全时间尺度上的服务器占用数量。对比3种算法在各个时刻求解的决策变量$\boldsymbol{X}(t)$，本文提出的DG Alg.算法能够在98.62%的时刻求得最优解而Greedy算法仅能在88.54%的时刻取得最优解。图11通过MATLAB仿真对比了3种算法在不同规格的网络请求下运行所需的CPU时间，仿真实验中选取VNF资源需求数

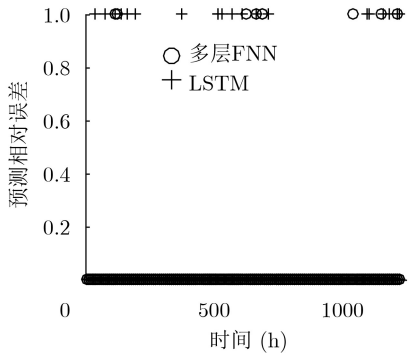

图 7 小规格VNF实例预测误差

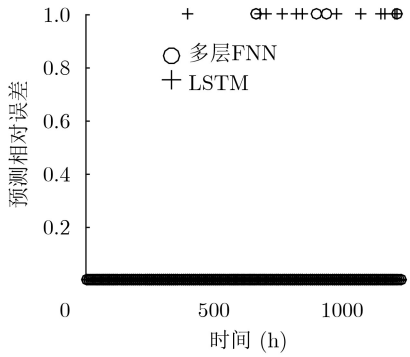

图 8 中规格VNF实例预测误差

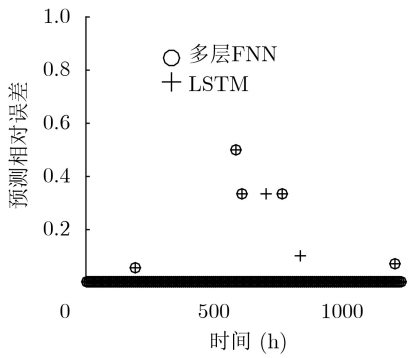

图 9 大规格VNF实例预测误差

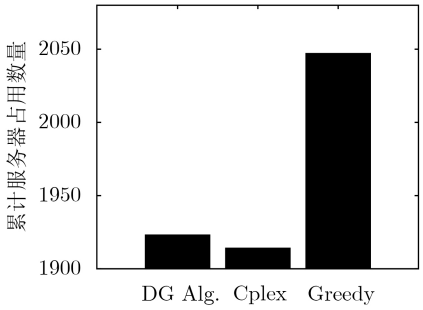

图 10 累计服务器占用数量对比

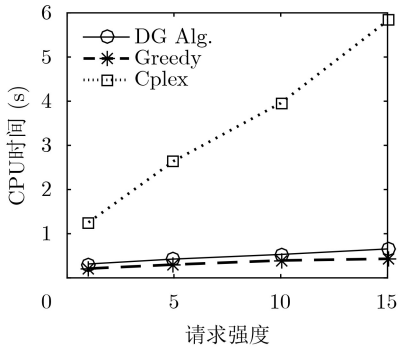

图 11 算法CPU运行时间对比

据集中的历史请求数据作为单位请求强度。通过对比，本文提出的DG Alg.算法表现出与Greedy算法相近的时间复杂度，而求解理论最优值的Cplex方法时间复杂度明显高于其他两种算法，当请求的VNF实例数量超过1000个后，Cplex无法继续计算其理论最优值。

6　结束语

本文提出了一种自适应的VNF资源容量调整方法。该方法结合已有的LSTM流量预测方法提出了一种基于多层FNN的VNF资源需求预测方法，降低了流量预测误差对资源需求预测的影响，实现了VNF资源需求的精确预测。然后利用二次分配模型建模VNF动态部署问题，并设计了动态编码遗传算法求解，实现了VNF实例的集中部署。后续工作将研究如何降低VNF资源需求预测误差导致的额外部署开销，进一步提高VNF动态部署的资源利用效率。

参 考 文 献

[1] ORDONEZ-LUCENA J, AMEIGEIRAS P, LOPEZ D, *et al*. Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges[J]. *IEEE Communications Magazine*, 2017, 55(5): 80–87. doi: 10.1109/mcom.2017.1600935.

[2] HERRERA J G and BOTERO J F. Resource allocation in NFV: A comprehensive survey[J]. *IEEE Transactions on*

Network and Service Management, 2016, 13(3): 518–532. doi: 10.1109/tnsm.2016.2598420.

[3] ADAMUZ-HINOJOSA O, ORDONEZ-LUCENA J, AMEIGEIRAS P, *et al*. Automated network service scaling in NFV: Concepts, mechanisms and scaling workflow[J]. *IEEE Communications Magazine*, 2018, 56(7): 162–169. doi: 10.1109/mcom.2018.1701336.

[4] RAHMAN S, AHMED T, HUYNH M, *et al*. Auto-scaling VNFs using machine learning to improve QoS and reduce cost[C]. Proceedings of 2018 IEEE International Conference on Communications, Kansas City, USA, 2018: 1–6.

[5] TANG Hong, ZHOU D, and CHEN Duan. Dynamic network function instance scaling based on traffic forecasting and VNF placement in operator data centers[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2018, 30(3): 530–543. doi: 10.1109/tpds.2018.2867587.

[6] ALAWE I, HADJADJ-AOUL Y, KSENTINI A, *et al*. Smart scaling of the 5G core network: An RNN-based approach[C]. Proceedings of 2018 IEEE Global Communications Conference, Abu Dhabi, The United Arab Emirates, 2018: 1–6.

[7] ALAWE I, HADJADJ-AOUL Y, KSENTINIT A, *et al*. An efficient and lightweight load forecasting for proactive scaling in 5G mobile networks[C]. Proceedings of 2018 IEEE Conference on Standards for Communications and Networking, Paris, France, 2018: 1–6.

[8] FEI Xincai, LIU Fangming, XU Hong, *et al*. Adaptive VNF scaling and flow routing with proactive demand prediction[C]. IEEE Conference on Computer Communications, Honolulu USA, 2018: 486–494.

[9] 唐伦, 周钰, 杨友超, 等. 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[J]. 电子与信息学报, 2019, 41(9): 2071–2078. doi: 10.11999/JEIT180894.
TANG Lun, ZHOU Yu, YANG Youchao, *et al*. Virtual network function dynamic deployment algorithm based on prediction for 5G network slicing[J]. *Journal of Electronics & Information Technology*, 2019, 41(9): 2071–2078. doi: 10.11999/JEIT180894.

[10] REN Yi, PHUNG-DUC T, LIU Yikuan, *et al*. ASA: Adaptive VNF scaling algorithm for 5G mobile networks[C]. Proceedings of 2018 IEEE 7th International Conference on Cloud Networking, Tokyo, Japan, 2018: 1–4.

[11] WANG Xiaoke, WU Chuan, LE F, *et al*. Online VNF scaling in datacenters[C]. Proceedings of 2016 IEEE 9th International Conference on Cloud Computing, San Francisco, USA, 2016: 140–147.

[12] WANG Xiaoke, WU Chuan, LE F, *et al*. Online learning-assisted VNF service chain scaling with network uncertainties[C]. Proceedings of 2017 IEEE 10th International Conference on Cloud Computing, Honolulu, USA, 2017: 205–213.

[13] 史久根, 张径, 徐皓, 等. 一种面向运营成本优化的虚拟网络功能部署和路由分配策略[J]. 电子与信息学报, 2019, 41(4): 973–979. doi: 10.11999/JEIT180522.
SHI Jiugen, ZHANG Jing, XU Hao, *et al*. Joint optimization of virtualized network function placement and routing allocation for operational expenditure[J]. *Journal of Electronics & Information Technology*, 2019, 41(4): 973–979. doi: 10.11999/JEIT180522.

[14] 张红旗, 黄睿, 常德显. 一种基于匹配博弈的服务链协同映射方法[J]. 电子与信息学报, 2019, 41(2): 385–393. doi: 10.11999/JEIT180385.
ZHANG Hongqi, HUANG Rui, and CHANG Dexian. A collaborative mapping method for service chain based on matching game[J]. *Journal of Electronics & Information Technology*, 2019, 41(2): 385–393. doi: 10.11999/JEIT180385.

[15] ZHOU Songnian. A trace-driven simulation study of dynamic load balancing[J]. *IEEE Transactions on Software Engineering*, 1988, 14(9): 1327–1341. doi: 10.1109/32.6176.

[16] MEDHAT A M, TALEB T, ELMANGOUSH A, *et al*. Service function chaining in next generation networks: State of the art and research challenges[J]. *IEEE Communications Magazine*, 2017, 55(2): 216–223. doi: 10.1109/mcom.2016.1600219rp.

[17] PARVEZ I, RAHMATI A, GUVENC I, *et al*. A survey on low latency towards 5G: RAN, core network and caching solutions[J]. *IEEE Communications Surveys & Tutorials*, 2018, 20(4): 3098–3130. doi: 10.1109/comst.2018.2841349.

[18] GOODFELLOW I, BENGIO Y, and COURVILLE A. Deep Learning[M]. Cambridge, USA: MIT Press, 2016: 397–399.

[19] HORNIK K. Approximation capabilities of multilayer feedforward networks[J]. *Neural Networks*, 1991, 4(2): 251–257. doi: 10.1016/0893-6080(91)90009-t.

[20] LOIOLA E M, DE ABREU N M M, BOAVENTURA-NETTO P O, *et al*. A survey for the quadratic assignment problem[J]. *European Journal of Operational Research*, 2007, 176(2): 657–690. doi: 10.1016/j.ejor.2005.09.032.

袁 泉：男，1991年生，博士生，研究方向为移动核心网体系架构、网络功能虚拟化.

游 伟：男，1984年生，讲师，研究方向为密码学、5G网络安全.

季新生：男，1968年生，教授、博士生导师，研究方向为5G网络安全、移动通信网络体系架构.

汤红波：男，1968年生，教授、博士生导师，研究方向为5G网络安全、移动通信网络体系架构.

责任编辑：陈 倩