面向云服务提供商的服务选择方法研究

张云勇[1]，李素粉[1]，吴俊[2]，房秉毅[1]
（1. 中国联通研究院，北京 100048；2. 北京邮电大学 经济管理学院，北京 100876）

摘 要：从云服务提供商角度出发，为提高服务选择的有效性，首先对服务选择流程进行分析，提出面向云服务提供商的服务选择思路，设计服务选择具体流程。进而从服务提供商角度出发，分析影响服务选择结果的主要因素，建立基于用户需求偏好和服务资源调度的服务选择数学模型，设计智能算法进行求解。最后进行算例分析，实验结果验证了该方法的可行性和有效性。

关键词：云计算；服务选择；服务质量；服务调度
中图分类号：TP393 **文献标识码：**A **文章编号：**1000-436X(2012)09-0066-11

Research on the cloud services provider-oriented services selection method

ZHANG Yun-yong[1], LI Su-fen[1], WU Jun[2], FANG Bing-yi[1]
(1. China Unicom Research Institute, Beijing 100048, China;
2. School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: In order to select services more accurately, firstly, a detailed process for services selection was proposed from the perspective of cloud services providers. Secondly, a services selection mathematical programming model was proposed for the selection of cloud services by considering two kinds of factors: QoS attributes and services scheduling related parameters, and the model was solved by the GA-based algorithm. Finally, a case study is conducted to demonstrate the feasibility and effectiveness of the proposed approach.

Key words: cloud computing; services selection; quality of service; service scheduling

1 引言

云计算作为信息通信技术（ICT, information and communication technology）转型机遇的发展趋势，已经得到了国内外业界的广泛认同，很多市场调研机构纷纷预测云服务将成为全球增长速度最快的市场，我国政府在国家战略高度也积极推进国家云计划。

云计算环境下，一些大型企业纷纷将传统业务向云服务转型，组建企业云，形成云服务资源池。例如，亚马逊建立专门的数据中心，以云服务的形式向用户和开发者提供存储和计算资源。云计算发展趋势下，企业的业务以云服务形式进行共享和集成。随着云计算在企业的广泛应用，必将催生大量服务。以电信行业为例，国内外电信运营企业纷纷开展云计算相关研究和应用，逐渐形成基础设施即服务（IaaS, infrastructure as a service）、平台即服务（PaaS, platform as a service）和软件即服务（SaaS, software as a service）3 个层次的服务资源池，服务资源将呈现规模化和商业化特点。这种情况下，对

收稿日期：2012-01-31；修回日期：2012-05-16
基金项目：国家自然科学基金资助项目（71172134）
Foundation Item: The National Natural Science Foundation of China (71172134)

于企业内部或外部用户需求，有效的服务选择方法，对于提高服务选择结果的有效性和服务资源的整体利用率具有重要意义。

随着服务计算思想在企业界和学术界的扩展与渗透，服务选择相关研究近几年得到国内外学者重视，取得了丰富的研究成果。服务计算环境下的服务选择是指根据用户需求，包括功能性需求和非功能性需求，从服务资源池中选择出满足用户需求的服务。

基于功能需求的服务匹配是指根据用户对服务的功能性需求描述，例如服务的输入输出接口描述，从服务资源池内选择出与功能需求相匹配的服务流程，服务流程由一个或者多个抽象服务节点依一定的流程结构组合而成，每个抽象服务对应服务资源池内一个或者多个功能相同的具体服务。相关研究主要是针对 Web 服务选择展开，采用基于关键词[1]和基于语义的服务[2~4]匹配方法等。查全率和查准率是关键词匹配方法面临的两大研究难点，基于语义的方法可以改善这一问题。语义 Web 服务的语义信息是以本体为基础实现，而现实应用系统中难以构建完善的本体库，且语义异构等现象普遍存在，信息描述不能对语义逻辑提供充分的支持。文献[4]通过定义良好的语义信息，提高服务选择方法的效率。

基于非功能性需求的服务选择是指基于特定的抽象服务流程，根据用户对服务的非功能性需求以及服务资源池内服务的运行状态等信息，为抽象服务流程中的每个节点选择出一个或者一组具体服务。相关研究主要集中在基于 QoS 需求的服务选择方法研究，包括服务质量模型[5,6]、QoS 局部最优服务选择[7]、QoS 全局最优服务选择[8~10]以及基于信任[11,12]和服务关联[13,14]等因素的服务选择。

上述研究主要是针对服务计算环境下服务选择的特点展开。服务计算环境下，服务资源具有海量特点，同时服务类型多种多样，也具有海量特点，各服务提供商通过公共注册中心进行服务注册，由服务中介或者服务选择平台根据用户需求进行服务选择。因此，服务计算环境下服务选择具有 2 个特点：面向海量服务资源和面向用户。基于这 2 个特点，现有相关研究主要集中在：基于用户需求（包括对服务的功能需求和非功能需求），在海量服务资源中选择出最优[8]或者满足用户需求[15]的服务。

云计算环境下服务选择在一定程度上与服务计算环境下的服务选择类似，但同时具有自身的特点。随着云计算在企业的广泛应用，必将催生大量云服务，这一点与服务计算环境类似，可以借鉴相关的服务选择方法和研究成果。然而，在同一朵云中，服务的类型一般不会出现海量特点，例如，亚马逊云，提供的服务类型包括存储、计算、数据库和网络等 11 类[16]，其中以计算和存储服务为主。云计算环境下，政府或者大型集团企业（在本文中统称为服务提供商）构建“政府云”或“企业云”，服务提供商根据用户需求进行服务选择。服务选择过程不仅考虑用户需求，同时关注提供商服务资源池的资源调度问题。因此，与服务计算环境下服务选择相比，云计算环境下服务资源选择具有 2 个特点：一方面，服务资源类型有限；另一方面，在考虑用户需求的同时兼顾服务资源调度。目前，服务选择相关研究主要集中在基于功能和非功能需求的服务选择方法，兼顾服务资源调度的服务选择相关研究鲜有提及。

本文基于服务选择现有研究基础，针对云计算环境下的服务选择特点，从服务提供商角度出发，研究云计算环境下的服务选择问题。首先，分析云服务提供商的服务资源特点，对服务资源池进行形式化描述。进而，提出一种兼顾用户需求和服务资源调度的服务选择方法，重点分析 3 种服务资源调度原则，分别构建计算规则，并将其引入到服务选择模型，建立基于用户需求和服务资源调度的服务选择数学模型。设计智能算法进行求解。最后进行算例分析。

2　面向云服务提供商的服务选择流程

为清楚地描述服务选择流程，首先对云服务资源池进行形式化描述。然后，从云服务提供商角度出发，给出服务选择总体思路，并设计具体服务选择流程。

2.1　云服务资源池形式化描述

云服务资源池由多个云服务构成，每个云服务具有一定的业务类型。为便于描述，首先对云服务资源池的划分进行假设并对相关概念进行形式化定义。

假设 1　假设云服务提供商的服务资源能够按照功能进行分组，将所有服务资源分为若干个服务组，记为 *ServiceSet*，形成服务组集合，记为 *ServiceSetS*，可以描述为式（1）。

$$ServiceSetS=\{ServiceSet_1, ServiceSet_2, ServiceSet_3,\cdots,ServiceSet_n,\cdots,ServiceSet_N\} \quad (1)$$

其中，N 为自然数。

根据假设 1，同一个服务组中的服务具有相同或相似的业务功能，不同服务组中的服务具有不同的业务功能。每个服务组由一个抽象服务标识，描述为

$$ServiceSet=(GroupID, FuncSet) \quad (2)$$

其中，*GroupID* 是服务组的唯一性标识；*FuncSet* 是服务组 *ServiceSet* 内云服务的业务功能描述，本文用三元组进行描述，如式(3)。

$$FuncSet=(Function, Input, Output) \quad (3)$$

其中，*Function* 是服务的业务功能描述；*Input* 和 *Output* 分别是输入和输出接口描述。

定义 1 云服务指云服务资源池内具有特定业务功能和非功能描述的具体服务，本文用六元组对云服务进行形式化描述，如式（4）。

$$S=(GroupID, ID, FuncSet, QoSSet, Provider, Site) \quad (4)$$

其中，S 表示云服务；*GroupID* 指服务 S 所属的服务组编号；*ID* 是服务 S 在服务组 *GroupID* 中的编号，*GroupID* 与 *ID* 一起形成云服务 S 的唯一性标识；*FuncSet* 是服务 S 的功能描述，定义同式（3）；*QoSSet* 是服务 S 的质量描述，*QoSSet* 是一个复合参数，用参数向量表示 $QoSSet=(QoS_1, QoS_2, QoS_3, \cdots)$，一般包括服务的执行时间、费用和可靠性等指标，不同类型服务需要的 QoS 指标可能不同，随应用场景变化，服务的 *QoSSet* 参数值会发生变化；*Provider* 指服务 S 的提供单位；*Site* 指服务所在的物理位置。

对于云服务资源池内的云服务，提供模式一般分为 2 种：1）单一服务提供，根据用户需求，选择资源池内的一个服务提供给用户；2）组合服务提供，根据用户需求，选择资源池内的多个服务形成服务流程，提供给用户，例如亚马逊多个云服务的结合使用[16]，如图 1 所示。

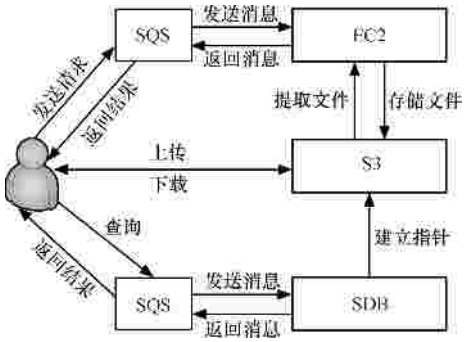

图 1 亚马逊多个云服务的结合使用

鉴于云服务提供商的云服务类型有限，例如亚马逊目前提供 11 类云服务[16]，容易挖掘出潜在组合服务流程。可以依据历史应用信息和服务之间的业务关联进行服务流程挖掘。服务流程挖掘不是本文研究的重点，这里不再详细讨论。在实例化之前，组合服务流程是一类抽象服务，本文用二元组对其进行描述，如式(5)所示。

$$ServiceFlow=(FlowID, FlowFuncSet) \quad (5)$$

所有的云服务流程形成云服务流程集合，记为 *ServiceFlowS*。

抽象云服务集合 *ServiceSetS* 和抽象云服务流程集合 *ServiceFlowS* 一起描述了服务资源池能够提供的所有服务类型。本文采用五元组对服务资源池进行形式化描述，如式(6)。

$$ServicePool=(PoolID, PoolName, ServiceSetS, ServiceFlowS, PoolOwner) \quad (6)$$

其中，*PoolID*、*PoolName* 和 *PoolOwner* 分别描述服务资源池的唯一标识、名称和所属企业等基本信息；*ServiceSetS* 和 *ServiceFlowS* 的含义分别见式（1）和式（5）。

2.2 面向提供商的云服务选择流程

面向云服务提供商的服务选择总体思路可以描述为：面对用户的服务请求，首先分析用户的服务需求，将其抽象为功能需求集合和非功能需求集合，作为服务选择的依据，这里的非功能需求用服务质量（QoS）描述。然后根据功能需求集合依次在服务组集合 *ServiceSetS* 和服务流程集合 *ServiceFlowS* 中进行功能匹配，找到满足功能需求的服务，形成抽象服务描述 *ServiceSet* 或抽象服务流程描述 *ServiceFlow*，每个抽象服务对应多个功能相同或相似的具体服务 S。进而，综合考虑用户的 QoS 需求和服务资源调度，为每一个抽象服务选择一个具体服务。最后形成服务选择结果，提供给用户。

基于上述思路，面向云服务提供商的服务选择具体流程如图 2 所示，具体描述如下。

步骤 1 根据给定的用户需求，用户需求包括对服务的功能需求描述和非功能需求描述，云服务选择系统根据需求描述，将用户需求抽象为用户功能需求集合和非功能需求集合，分别用符号 *UserFuncSet* 和 *UserQoSSet* 描述。

步骤 2 服务选择系统根据 *UserFuncSet* 中的参数需求，在服务资源池中查找功能匹配的服务组。匹配方法如下。

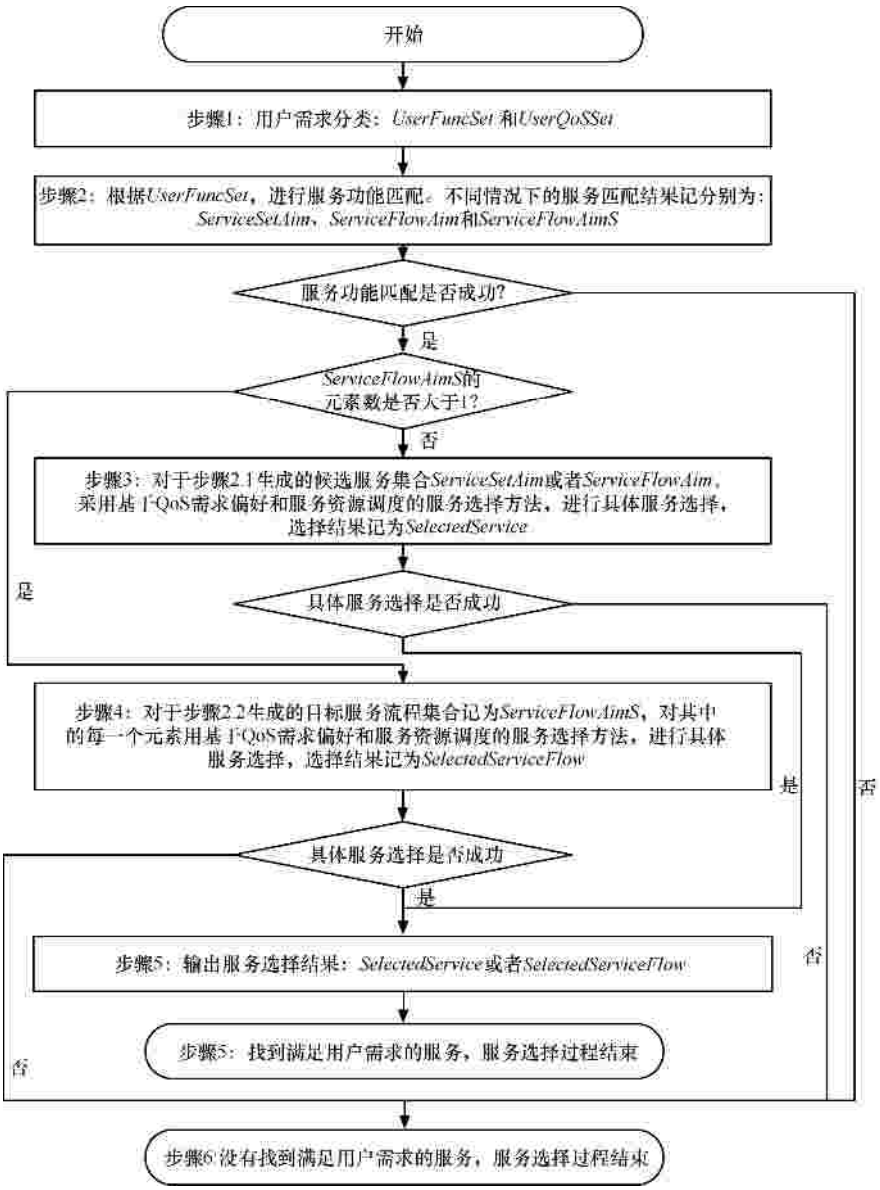

图 2　面向云服务提供商的服务选择流程

步骤 2.1　进行单一服务匹配，根据用户需求 *UserFuncSet* **参数在服务组集合** *ServiceSetS* **中对抽象服务** *ServiceSet* **进行匹配，查找** *FuncSet* **各参数均与用户需求相应参数匹配的服务组** *ServiceSet*。

根据假设 1，**匹配结果有** 2 **种：**0 **或** 1。**若返回结果为** 1，**则表明服务资源池中存在满足用户功能需求的单一服务，将该服务所在的服务组** *ServiceSet* **称作候选服务集合，记为** *ServiceSetAim*，**转步骤** 3；**若返回结果为** 0，**则表明服务资源池中的任一服务均不能单独实现用户的功能需求，转步骤** 2.2。

步骤 2.2　进行服务流程匹配，根据用户需求 *UserFuncSet* **参数在抽象云服务流程集合** *ServiceFlowS* **中对** *ServiceFlow* **进行匹配，查找** *FlowFuncSet* **各参数均与用户需求相应参数匹配的服务流程** *ServiceFlow*。**匹配结果有** 3 **种情况：**1)0；2)1；3) **大于** 1。**对于情况** 1)，**转步骤** 6。**对于情况** 2)，**将匹配的服务流程** *ServiceFlow* **称作目标服务流程，记为** *ServiceFlowAim*，**转步骤** 3。**对于情况** 3)，**将匹配的多个服务流程** *ServiceFlow* **称作目标服务流程集合，记为** *ServiceFlowAimS*，**转步骤** 4。

步骤 3　对于步骤 2.1 **生成的候选服务集合** *ServiceSetAim*，**或者步骤** 2.2 **生成的目标服务流程** *ServiceFlowAim*，**分析用户的** QoS **需求参数** *UserQoSSet*，**采用基于** QoS **需求偏好和服务资源调**

度的服务选择方法（见本文第 3 节），进行具体服务选择。服务选择结果有 2 种：成功或失败。成功则输出选择结果，记为 *SelectedService*，转步骤 5。失败则转步骤 6。

步骤 4 对于步骤 2.2 生成的目标服务流程集合 *ServiceFlowAimS*，对其中的每一个元素 *ServiceFlow* 用基于 QoS 需求偏好和服务资源调度的服务选择方法（见本文第 3 节），进行服务选择。同样，服务选择结果有 2 种：成功或失败。对于成功服务流程 *ServiceFlow*，进行标记和统计，数量记为 *SFN*。若 *SFN*=0，转步骤 6；若 *SFN*=1，则输出服务选择结果，转步骤 5；若 *SFN*>1，说明存在具体服务流程满足用户需求，则从促进服务资源的均衡利用角度出发，以最优资源调度为原则，选择其中一个具体服务流程，作为服务选择结果，记为 *SelectedServiceFlow*。转步骤 5。

步骤 5 找到满足用户需求的服务资源，选择的服务资源为 *SelectedService* 或者 *SelectedServiceFlow*，服务选择过程结束。

步骤 6 没有找到满足用户需求的服务，服务选择过程结束。

3 基于 QoS 需求偏好和服务资源调度的服务选择方法

针对第 2 节服务选择流程中步骤 3 和步骤 4，本节给出一种基于 QoS 需求偏好和服务资源调度的服务选择方法。

3.1 符号定义

为便于建立基于数学模型，首先给出表 1 所示符号定义。

3.2 模型要素分析

基于 QoS 需求偏好和服务资源调度的服务选

表 1 基于 QoS 需求偏好和服务资源调度的服务选择模型基本符号

符号	意义及说明
i	组合服务流程 $SelectedServiceFlowrtot$, $r=1,2,3,\cdots,R$, $t=1,2,3,\cdots,T$ 中每个服务节点的序号，$i=1,2,3,\cdots,I$，I 是正整数
j	每个 *ServiceSet* 内服务的 *ID*，$j=1,2,3,\text{L},J_i$，J_i 是正整数，$i=1,2,3,\cdots,I$
$S_{i,j}$	服务组 i 中的服务 j
$Y_{i,j}$	0-1 变量，决策变量 $j=1,2,3,\text{L},J_i, i=1,2,3,\text{L},I$ 当服务 $S_{i,j}$, $j=1,2,3,\text{L},J_i, i=1,2,3,\text{L},I$ 被选择时，$Y_{i,j}=1$，否则 $Y_{i,j}=0$
$q_{i,j}^{k}$	服务 $S_{i,j}$, $j=1,2,3,\text{L},J_i, i=1,2,3,\text{L},I$ 的第 k 个 QoS 指标
$F1,F21,F22,F3$	依次分别表示资源集约利用惩罚函数、不饱和运行状态下的资源均衡利用惩罚函数、饱和运行状态下的资源均衡利用惩罚函数和物理距离惩罚函数
$P1,P21,P22,P3$	惩罚系数，$P1\geq 0, P21\geq 0, P22\geq 0, P3\geq 0$
$ServerStatus$	物理服务器状态标识，$ServerStatus=0$ 表示处于开启状态；$ServerStatus=1$ 表示处于关闭状态
$ServerUtilizRatio$	物理服务器的状态参数，表示当前虚机利用率，$0\leq ServerUtilizRatio\leq 1$
$pn_{i,j}$	服务 $S_{i,j}$ 当前并行实例数，$pn_{i,j}=0,1,2,3,\text{L},pN_{i,j}$
$pN_{i,j}$	服务 $S_{i,j}$ 允许的最大并行实例数，$pN_{i,j}$ 是正整数
$st_{i,j,1}$	$st_{i,j,1}\in\{0,1\}$，其中，$st_{i,j,1}=0$，表示服务 $S_{i,j}$ 处于不饱和工作状态；$st_{i,j,1}=1$，表示服务 $S_{i,j}$ 处于饱和状态，$i=1,2,3,\text{L},I, j=1,2,3,\text{L},J_i$
$st_{i,j,2}$	$st_{i,j,2}\in\{0,1\}$，表示服务 $S_{i,j}$ 上进度最快的实例的执行进度，$i=1,2,3,\text{L},I, j=1,2,3,\text{L},J_i$
$queue_{i,j}$	服务 $S_{i,j}$ 处于饱和运行状态时，等待服务 $S_{i,j}$ 进行实例化的请求数量
$t_{i,j}$	服务 $S_{i,j}$ 进行一次实例化所需要的时间
$wt'_{i,j}$	服务 $S_{i,j}$, $j=1,2,3,\text{L},J_i, i=1,2,3,\text{L},I$ 的排队等待时间，即从选择该服务到可以被当前用户调用的时间
$wt''_{i,j}$	对于服务功能匹配结果是组合服务流程的情况，流程中服务 $S_{i,j}$, $j=1,2,3,\text{L},J_i, i=1,2,3,\text{L},I$ 的任务到达时间
$wt_{i,j}$	对于当前用户，若选择 $S_{i,j}$ 需要的实际等待时间，由服务等待时间 $wt'_{i,j}$ 和任务到达时间 $wt''_{i,j}$ 决定
Q^k	组合服务流程的第 k 个 QoS 指标
Q_0^k	对于服务选择结果的第 k 个 QoS 指标，用户给定的限值

择模型的要素可以分为 2 类：QoS 指标和服务资源调度相关指标。本节对这 2 类指标进行分析，并给出各指标的计算规则。

3.2.1 QoS 指标

QoS 指标一般包括服务的执行时间、费用、可用性、可靠性和处理能力等。云服务分为 3 个层次 SaaS、PaaS 和 IaaS，对于不同层次的服务。本文用向量 $Q=(Q^1, Q^2, Q^3, \cdots, Q^k, \cdots, Q^K)$ 描述用户对服务的 QoS 指标需求，Q^1 描述用户最关心的指标，Q^2 次之，以此类推。服务的 QoS 参数值会出现在服务描述中。因此，各个服务的 QoS 指标相关参数值可以从服务描述中获取。

3.2.2 服务资源调度相关指标

云服务资源调度是指基于调度规则对云服务资源池中的云服务进行合理有效的调节和利用。本文主要考虑 3 种调度原则：资源集约利用原则、资源均衡利用原则和物理距离最短原则。本文第 2.2 节的服务选择流程，基于功能的服务匹配结果有 2 种：单服务和服务流程。为简化描述计算过程，将第 1 种结果（单服务）看作仅包含一个服务节点的服务流程。

1) 资源集约利用原则

该调度原则主要针对物理服务器的资源利用效率问题。目前，服务器资源是云服务的一个主要内容，一般通过虚拟化技术将物理服务器虚拟为多个独立虚拟服务器（简称为虚机），将虚机作为独立的云服务器资源向用户提供。为提高物理服务器资源的利用率，本文设计了资源集约利用原则，核心思想是优先选择处于开启状态的物理服务器上的虚拟服务器资源。这一原则在服务选择模型中将以资源集约利用函数 $F1$ 体现，$F1$ 是物理服务器开启状态的函数。

$F1$ 计算思路为：判断服务 $S_{i,j}$ 所在物理服务器状态参数 $ServerStatus_{i,j}$，$ServerStatus_{i,j}=0$ 表示处于开启状态；$ServerStatus_{i,j}=1$ 表示处于关闭状态。$F1$ 的具体计算规则可以描述为式(7)。

$$F1=\sum_{i=1}^{I}\sum_{j=1}^{J_i}\left(X_{i,j}Y_{i,j}ServerStatus_{i,j}\right) \quad (7)$$

其中，$X_{i,j}\in\{0,1\}, i=1,2,3,\cdots,I, j=1,2,3,\cdots,J_i$，当服务 $S_{i,j}$ 是虚机资源时，$X_{i,j}=1$；否则 $X_{i,j}=0$。

2) 资源均衡利用原则

本文从 2 个方面考虑资源均衡利用原则：不饱和运行状态下的云服务资源均衡利用原则和饱和状态下云服务资源均衡利用原则。

不饱和运行状态下的云服务资源均衡利用原则，主要针对虚机（一类云服务资源）的选择问题进行分析，对于已开启的物理服务器资源，如何均衡利用其上的虚机资源。也就是说，当存在多个同类虚机可供选择，这些虚机均处于空闲状态且所在物理服务器为开启状态时，怎样选择其中一个虚机需要遵循一定的原则。本文采用的实现规则是优先选择当前负荷较小的物理服务器上的虚机资源。这一原则在服务选择模型中将以函数 $F21$ 体现，$F21$ 是物理服务器的虚机利用率函数。$F21$ 的计算思路是：获取候选服务 $S_{i,j}$ 所在物理服务器当前时刻的虚机利用率 $ServerUtilizRatio_{i,j}$（已使用的虚机数/总虚机数），选择利用率最小的物理服务器上的虚机资源。$F21$ 的具体计算规则可以描述为式（8）。

$$F21=\sum_{i=1}^{I}\sum_{j=1}^{J_i}\left(X_{i,j}Y_{i,j}(1-ServerStatus_{i,j})\times ServerUtilizRatio_{i,j}\right) \quad (8)$$

其中，$X_{i,j}$ 和 $ServerStatus_{i,j}$ 的含义同式（7）。

饱和状态下云服务资源均衡利用原则，主要针对候选服务资源均处于忙碌状态时的服务选择问题进行分析。本文采用的实现规则是优先选择等待时间最短的服务资源。这一原则在服务选择模型中将函数 $F22$ 体现，$F22$ 是服务资源的实际等待时间函数。

在给出 $F22$ 的计算规则之前，首先分析各候选服务的等待时间计算方法。

不失一般性，对服务等待进行假设，如假设 2。

假设 2　对于服务 $S_{i,j}$，如果当前运行实例数 $pn_{i,j}$ 等于最大并行实例数 $pN_{i,j}$，则新的实例请求需要排队等待。

服务的运行状态包括：不饱和运行状态（运行实例数小于 $pN_{i,j}$）、饱和运行状态（运行实例数等于 $pN_{i,j}$）和排队等待个数，由系统动态获取。如果服务的运行实例数小于 $pN_{i,j}$，则等待时间为零；若运行实例数等于 $pN_{i,j}$，则根据服务的执行时间和执行进度计算等待时间；若服务有排队，则根据其当前执行进度、执行时间、最大并行实例数 $pN_{i,j}$ 和排队数量计算等待时间。基于此，服务 $S_{i,j}$ 等待时间

的计算规则可以描述为式（9）。

$$wt'_{i,j} = st_{i,j,1} \times \left\{ \mathrm{int}\left(\frac{queue_{i,j}}{pN_{i,j}} \right) + \left(1 - st_{i,j,2}\right) \right\} \times t_{i,j} \quad (9)$$

其中，int( ) 表示取整函数，这里采用上取整规则。其他各参数定义见表 1。

对于一个组合服务流程，选择具体服务之后，一般不会立刻使用所有服务。例如对于流程中任一服务（开始服务节点除外），只有等前一个串联服务执行结束后，即任务到达后，才会调用其选择的具体服务资源。也就是说，一方面服务资源被释放后才可进行实例化，另一方面任务到达后才会使用服务资源。因此，本文综合考虑服务等待时间和任务到达时间，采用如下迭代计算过程计算流程中各个服务的实际等待时间：

Step1：For $i=1$

$$wt_{1,j} = wt'_{1,j},\ j = 1,2,3,\mathrm{L},J_i \quad (10)$$

$$wt''_{1,j} = 0,\ \ j = 1,2,3,\mathrm{L},J_i \quad (11)$$

Step2：For $i=2,3,4,\mathrm{L},I$

$$wt''_{i,j} = TotalTf\left(wt''_{k,j} + t_{k,j}, k = 1,2,3,\mathrm{L},i-1\right),$$
$$j = 1,2,3,\mathrm{L},J_i \quad (12)$$

$$wt_{i,j} = \max\left(wt'_{i,j} - wt''_{i,j}, 0\right),\ \ j = 1,2,3,\mathrm{L},J_i \quad (13)$$

其中，$TotalTf$ ( ) 是服务流程总体时间类参数的计算函数，具体计算公式参见文献[15]，$TotalTf\left(wt''_{k,j} + t_{k,j}, k = 1,2,3,\mathrm{L},i-1\right)$ 表示服务流程中前 $i-1$ 个服务的执行时间与实际等待时间之和，也就是第 i 个服务 $S_{i,j}$ 的任务到达时间。

整个服务流程总的实际等待时间即为流程中每个服务的实际等待时间之和，$F22$ 的计算规则如式（14）。

$$F22 = TotalTf\left(\sum_{j=1}^{J_i} Y_{i,j} wt_{i,j}, i = 2,3,4,\mathrm{L},I\right) \quad (14)$$

其中，函数 $TotalTf$ ( ) 的含义与式（12）相同。

3) 物理距离最短原则

云中的服务资源分布在不同的物理位置。服务选择中，优先选择距离用户较近的服务资源。这一原则在服务选择模型中将以函数 $F3$ 体现，$F3$ 是物理距离的函数。服务流程的物理距离包括 2 个方面，一方面是每个服务与用户之间的物理距离，另一方面是两两服务之间的物理距离，这里主要考虑具有串联关系的两两服务。本文基于这 2 种距离构造函数 $F3$。

从服务 $S_{i,j}$ 描述中提取地理位置参数 $Site_{i,j}$，从用户需求中获取用户的位置参数，记为 $UserSite$。$F3$ 的计算规则如式（15）。

$$F3 = \begin{cases} Distf\left(\sum_{j=1}^{J_i} Y_{i,j} Site_{i,j}, UserSite\right) & I = 1 \\ \frac{1}{I}\sum_{i}^{I} Distf\left(\sum_{j=1}^{J_i} Y_{i,j} Site_{i,j}, UserSite\right) + \\ \sum_{i}^{I-1} Z_{i,j} Distf\left(\sum_{j=1}^{J_i} Y_{i,j} Site_{i,j}, \sum_{j=1}^{J_i} Y_{i,j} Site_{i+1,j}\right) & I > 1 \end{cases} \quad (15)$$

其中，$j = 2,3,4,\mathrm{L},J_i$；$Distf\left(a,b\right)$ 是距离函数，用于计算 a,b 两点之间的距离；$Z_{i,j} \in \{0,1\}$，$Z_{i,j} = 1$ 表示服务 $S_{i,j}$ 与 $S_{i+1,j}$ 之间是串联关系，否则 $Z_{i,j} = 0$。

3.2.3 总体指标的计算规则

对于需要选择组合服务流程的情况，需要计算组合服务的总体 QoS 指标。在不同的流程实例中，需要根据局部流程的结构特点来计算。流程的基本结构主要有 4 种：顺序、并行、选择和循环。本文第 2 节描述的服务选择流程，得到的服务组合流程将呈现上述 4 种或部分结构。关于 4 种不同流程结构下的总体指标计算方法，相关文献研究较多[8, 15]。本文中，总体指标的计算参见文献[15]的计算方法。

由于不同的指标具有不同的量纲，因此在将多个指标进行综合计算时，首先需要对各种指标进行无量纲归一化处理。本文采用极差规格化变换[17,18]对每一指标值进行无量纲归一化处理。无量纲归一化处理后，服务个体每一质量指标的取值都是[0, 1]区间上的无量纲点。本文所建模型中的参数均为归一化处理后的参数。

3.3 基于 QoS 需求偏好和服务资源调度的服务选择数学模型

基于 QoS 指标和服务的运行状态建立基于用户 QoS 需求偏好和服务资源调度的服务选择数学模型。基于 QoS 指标的服务选择是一个多目标数学规划问题，鉴于目标规划能较好的解决多目标决策问题，其目标函数不是寻求最大值或最小值，而是寻求这些目标与预计成果的最小差距，差距越小，目标实现的可能性越大。利用目标规划的这一特点，服务选择并不是以选择最优的服务为目标，而是找到满足用户需求的服务，在一定程度上可以避免优质服务排队现象，利于服务的均匀利用。

本文采用目标规划对基于 QoS 需求偏好和服务资源调度的服务选择问题进行建模。用户的 QoS 指标一般包括服务的执行时间、费用、可用性、可靠性和处理能力等。将所选服务或组合服务流程的实际 QoS 指标值与用户期望值的差值作为模型的目标项，差值越小表示越接近用户需求，当差值为 0 时，表示所选服务或组合服务流程完全满足用户需求。基于用户的 QoS 指标需求和服务的使用状态，目标规划的目标向量可以定义为：

目标向量是一个由多个目标项构成的一维向量，描述为式（16）。

$$Obj=(Obj_1,Obj_2,Obj_3,\cdots,Obj_o,\cdots,Obj_O) \tag{16}$$

其中，Obj_o 表示目标因子，$o=1,2,3,\cdots,O$，O 是自然数，$Obj_o \geq 0$。

在一次服务选择中，用户对于不同指标的偏好程度可能不同，例如：用户需求“在满足服务执行时间的条件下费用越低越好”，表明“执行时间”的优先等级高于“执行费用”。针对这一情况，提供一种灵活的目标项权重设置方法。该方法通过采用权重向量实现，目标权重向量定义如下：

目标权重向量是一个由多个权重因子构成的一维向量，描述为

$$ObjWeight=(a_1,a_2,a_3,\cdots,a_w,\cdots,a_W) \tag{17}$$

其中，a_w 表示权重因子，$w=1,2,3,\cdots,W$，W 是自然数；$0\leq a_w \leq 1$ 且 $\sum_{w=1}^{W} a_w=1$。

对于目标规划问题，将目标权重向量与目标向量的点乘乘积函数作为目标规划模型的目标函数，对于同一个目标函数，权重数量 W 与目标数量 O 相等，目标函数可以描述为

$$ObjFun=ObjWeight\cdot Obj^T=\sum_1^W a_w Obj_w \tag{18}$$

本文模型考虑 4 个等级的目标，前 3 个目标分别根据用户的 QoS 需求生成，QoS 指标需求一般可以分为 2 类，一类指标值越大越好，例如可用性、可靠性和处理能力等，在模型中用符号 $d_k^-, k=1,2,3$ 表示，另一类指标值越小越好，例如服务的执行时间和费用等，在模型中用符号 $d_k^+, k=1,2,3$ 表示；第 4 个目标是服务资源调度函数，根据这 4 个目标，形成目标向量 $Obj=(d_1^+V0, d_2^+V0, d_3^-V0, P1\times F1V0+P21\times F21V0+P22\times F22V0+P3\times F3V0)$，其中 4 个目标因子分别表示所选服务或组合服务流程的实际 QoS 指标值与用户期望值的差。根据用户需求偏好决定目标权重向量 $ObjWeight$。基于目标向量和权重向量，构造目标函数。模型的约束函数主要是目标约束和决策变量约束。

基于上述分析，建立基于 QoS 需求偏好和服务资源调度的服务选择数学模型，模型描述如式(19)~式(22)。

$$ObjFunc=a_1(d_1^+V0)+a_2(d_2^+V0)+a_3(d_3^-V0)+a_4(P1\times F1V0+P21\times F21V0+P22\times F22V0+P3\times F3V0) \tag{19}$$

s.t.

$$d_k^+=Q^k-Q_0^k,\ k=1,2 \tag{20}$$

$$d_3^-=Q_0^3-Q^3 \tag{21}$$

$$\sum_{j=1}^{J_i} Y_{i,j}=1,\ i=1,2,3,\cdots,I \tag{22}$$

其中，$I=1$ 表示单服务选择，$I>1$ 表示组合服务流程选择。式(18)为目标函数，表示服务或组合服务的 3 个 QoS 指标分别与用户期望值的差，依字典序分别达到最小，目标函数中的 V 表示“取大”操作，例如“d_1^+V0”表示“d_1^+”的值与“0”取大。式（20）、式（21）为模型的目标约束，表示 3 个 QoS 指标分别与用户期望值的差。式（22）是决策变量约束，表示为每个抽象服务选择一个具体服务，各参数的总体指标是 $Y_{i,j}$ 的函数。

3.4　模型求解

组合服务 QoS 局部和全局优化计算问题都是 NP 问题，算法具有指数复杂度，随着问题规模的增大，难以在多项式时间内找到问题的最优解[8,19]。遗传算法（GA, genetic algorithm）作为一种智能优化方法，具有并行计算、群体寻优的特点，已广泛应用于各种 NP-Complete 问题的求解[8]。鉴于此，本文基于遗传算法对所提服务选择方法进行仿真计算。

算法的基本思想为：基于第 2 节服务选择流程步骤 2，在服务资源池中找到目标服务组，形成组合服务流程（可以包含一个或多个服务节点）。将每一个配置后的组合服务流程编码为一个染色体，通过染色体之间的选择、交叉和变异等遗传操作，产生具有更高适应函数值的新染色体。这一过程不断重复进行，实现在解空间的并行全局搜索。算法停止时，得到一个染色体集合，对应模型的解集，也就是服务选择的方案集。

在遗传算法中，适应值是对染色体进行评价的

重要指标，适应值函数的构造非常重要。染色体对应问题的解，因此可以基于所求解问题的目标对染色体的适应值函数进行定义。

1) 适应值函数构建

根据模型的目标函数构建遗传算法的适应值函数 f'。

$$f' == a_1\left(d_1^+V0\right)+a_2\left(d_2^+V0\right)+a_3\left(d_3^-V0\right)+ a_4\left(P1\times F1V0+P21\times F21V0+ P22\times F22V0+P3\times F3V0\right) \tag{23}$$

用户对不同 QoS 指标的需求偏好通过 $a_k\,(k=1,2,3)$ 的取值来体现。

2) 适应值归一化处理

直观认识方面，适应值越高表示染色体性能越好，反之越差。因此便于直观理解，对适应值 f' 采用负指数方法进行归一化处理[15]，可以描述为

$$f=\mathrm{e}^{-mf'} \tag{24}$$

其中，$m>0$，是归一化参数。

归一化之后的适应值 f 是介于 0 和 1 之间的实数，$f=1$ 表示适应值最高，$f=0$ 表示适应值最低。

4 算例分析

4.1 参数初始化

本文基于云服务提供商的服务选择方法，服务选择结果可以是一个服务，也可以是由多个服务形成的组合服务流程。对于选择结果是一个服务的情况，本节算例分析中把其看作包含一个服务节点的服务流程，对本文第 3.3 节所建模型进行算例分析。服务流程包含 I 个服务节点，这里令 I=6，不失一般性，假设 I 个服务组的规模相同且规模 $J_i=20, i=1,2,3,\mathrm{L},I$。

1) 服务资源池中服务的 QoS 参数和资源调度相关参数

服务 $S_{i,j}$ 的相关参数值从服务描述中获取，例如：对于云计算 IaaS 层面的某存储服务 $S_{i,j}$，根据式(4)的云服务定义，有 i=*GroupID*，j=*ID* 且 *FuncSet* 和 *QoSSet* 分别描述该服务的功能和质量指标，*Provider* 和 *Site* 分别描述该服务的提供者信息和所处的物理位置信息。

这里 QoS 指标分别取服务执行费用、时间和可靠性。对于 QoS 指标参数和服务服务资源调度相关参数初始化，采用随机方法，在一定范围内自动生成，每个参数的取值范围设定如表 2 和表 3 所示，其中，$j=1,2,3,\mathrm{L},J_i, i=1,2,3,\mathrm{L},I$。

表 2 QoS 参数相关参数取值范围

参数	值
$q_{i,j}^1$	(0,10]
$q_{i,j}^2$	(0,10]
$q_{i,j}^3$	(0,1]

表 3 服务资源调度相关参数取值范围

参数	值
$ServerStatus_{i,j}$	{0,1}
$ServerUtilizRatio_{i,j}$	[0,1]
$st_{i,j,1}$	{0,1}
$st_{i,j,2}$	[0,1]
$queue_{i,j}$	(0,60]
$pN_{i,j}$	[1,20]
$pn_{i,j}$	[0, $pN_{i,j}$]

其中，参数 $queue_{i,j}$ 和 $pN_{i,j}$ 的取值是整数，参数 $queue_{i,j}$ 和 $pN_{i,j}$ 初始化方法分别是：在区间(0,60]和[1,20]上分别随机产生一个实数，然后取整分别作为参数 $queue_{i,j}$ 和 $pN_{i,j}$ 的值；参数 $pn_{i,j}$ 初始化方法是在区间[0, $pN_{i,j}$]上随机产生一个实数，然后取整作为该参数的值。

2) 用户需求相关参数设置

不同的用户需求一般体现在对各项指标的限值要求不同，本节实验假设了 3 种不同的用户需求情况，如表 4 所示。

表 4 用户需求相关参数设置

参数	Q_0^1	Q_0^2	Q_0^3	WT_0
Case 1	9	9	0.5	9
Case 2	8	8	0.5	8
Case 3	6	6	0.5	6

3) 目标权重参数

本文目标规划模型考虑了 4 个目标，对应的目标权重向量包含 4 个权重因子。为比较不同用户需求偏好情况下的服务选择结果，本算例对 4 种不同的目标权重向量取值下的服务选择进行计算分析，目标权重向量参数设置如表 5 所示。

表 5　目标权重参数设置

参数	(a_1,a_2,a_3,a_4)
ObjWeightA	(0.25,0.25,0.25,0.25)
ObjWeightB	(0. 75,0. 2,0.05,0)
ObjWeightC	(0. 5,0. 5,0,0)
ObjWeightD	(0,0 0,1)

其中，*ObjWeightA* 的 4 个权重取值相同，表示 4 个目标优先等级相同；*ObjWeightB* 和 *ObjWeightC* 的权重设置表示 3 个 QoS 指标具有从高到低不同的优先等级，未考虑资源调度问题；*ObjWeightD* 的权重设置表示只考虑资源调度目标而不考虑用户的 QoS 需求。

4) 遗传算法参数

遗传算法参数取交叉概率设为 0.6，变异概率设为 0.1，染色体种群规模取 25。

4.2　计算结果与分析

用 VC++6.0 编程实现求解算法。分别针对 3 种不同的用户 QoS 需求和 4 种不同的目标权重情况进行具体服务选择。通过设置不同的目标权重，决定在服务选择模型中是否引入资源调度规则，其中，*ObjWeightA* 和 *ObjWeightD* 权重设置下，在不同程度上考虑资源调度原则，*ObjWeightA* 和 *ObjWeightD* 权重设置下，不考虑资源调度原则。服务选择结果如表 6 和表 7 所示。

表 6　不同用户需求情况下的服务选择方案(*ObjWeightA*)

节点	1	2	3	4	5	6	适应值
Case1	8	13	10	2	10	13	1
Case2	11	17	20	12	6	9	1
Case3	16	6	10	11	15	13	0.75

表 7　不同目标权重参数情况下的服务选择方案(Case2)

节点	1	2	3	4	5	6	适应值
ObjWeightA	11	17	20	12	6	9	1
ObjWeightB	9	5	6	12	13	2	1
ObjWeightC	6	17	15	1	5	14	1
ObjWeightD	3	11	9	11	15	7	1

表 6 表明，3 种不同 QoS 需求情况下，前 2 种不同的用户需求情况下，模型的适应值都等于 1，意味着用户的需求都达到满意，但并一定是系统内服务选择的最优解或者最优方案。因此本文所提服务选择方法一方面能够满足用户的需求，另一方面在一定程度上能够减少优质服务排长队的问题。对于用户需求 Case3，算法运行结束时没有找到适应值等于 1 的服务流程，即完全满足需求的服务组合方案，这种情况下可以进而采用其他方法与用户达成协定，例如服务协商。

表 7 表明，4 种不同的目标权重参数下，都找到了满足需求的具体服务选择方案，而每种情况下的服务选择方案不尽相同。目标权重反映了用户对服务的 QoS 需求偏好和资源调度目标，表 6 的数据反映出不同的目标权重设置导致了不同的服务选择结果。其中，对于目标权重 *ObjWeightB* 和 *ObjWeightC*，参数设置为(0.75, 0.2, 0.05, 0)和(0.5, 0.5, 0, 0)，$a_4=0$ 表示服务选择计算过程中没有考虑资源调度因素，这种情况下的服务选择结果明显区别于其他 2 组，同时，由于 *ObjWeightB* 和 *ObjWeightC* 具有不同 QoS 权重，对应的服务选择结果也不相同。因此，将用户需求偏好和资源调度因素考虑到服务选择过程具有重要意义。

5　结束语

服务选择是云服务提供商进行服务提供时面临的一个首要问题。本文从云服务提供商角度，分析影响服务选择的主要因素，梳理服务选择流程，给出服务选择的具体实现步骤。进而建立基于用户需求偏好和服务资源调度的服务选择数学模型，设计智能算法进行求解。算例分析表明了方法的可行性和有效性。本文所提服务选择流程和模型对于提高云服务提供商的服务运营效率具有一定指导意义。

参考文献：

[1] 邓岳. 基于个性化服务匹配度的服务发现机制研究[D]. 西安 :西安电子科技大学 2007.
DENG Y. Research on Service Discovery Based on Personalized Service Matchmaking Degree[D]. Xi'an: Xidian University, 2007.

[2] YANG F C, SU S, LI Z. Hybrid QoS-aware semantic Web service composition strategies[J]. Science in China Series F: Information Sciences, Science in China Press, Co-published with Springer-Verlag GmbH, 2008, 51(11): 1822-1840.

[3] 白东伟. 基于语义的 Web 服务匹配与发现技术研究[D]. 北京: 北京邮电大学, 2008.
BAI D W. Research on Web Services Semantic Matchmaking and Discovery[D]. Beijing: Beijing University of Posts and Telecommunications, 2008.

[4] 狄小峰. 基于分布式本体的服务选择技术研究[D]. 北京: 清华大

学, 2012.
DI X F. Research on Service Selection Method based on Distributed Ontology[D]. Beijing: Tsinghua University, 2012.

[5] MENASCE D A. QoS issues in web services[J]. IEEE Internet Computing, 2002, 6: 72-75.

[6] RAN S. A model for web services discovery with QoS[J]. ACM, 2003, 4(1): 1-10.

[7] YU T, LIN K J. The design of QoS broker algorithms for QoS-capable web services[J]. International Journal of Web Service Research, 2004,1(4): 33-50.

[8] **刘书雷**, **刘云翔**, **张帆等**. **一种服务聚合中** QoS **全局最优服务动态选择算法**[J]. **软件学报**, 2007. 18(3): 646-656.
LIU S L, LIU Y X, ZHANG F, *et al*. A dynamic Web services selection algorithm with QoS global optimal in Web services composition [J]. Journal of Software, 2007, 18 (3): 646-656.

[9] FUDZEE M F M, ABAWAJY J H. QoS-based adaptation service selection broker[J]. Future Generation Computer Systems, 2011, 27: 256-264.

[10] GUO F, ZHAO L, WANG Y, *et al*. Research on the Web services selection problem[A]. Proceedings of the 2010 Second International Workshop on Education Technology and Computer Science (ETCS)[C]. Wuhan, China, 2010. 284 - 287.

[11] **潘静**, **徐锋**, **吕建**. **面向可信服务选取的基于声誉的推荐者发现方法**[J]. **软件学报**, 2010, 21(2): 388-400.
PAN J, XU F, LV J. Reputation-based recommender discovery approach for service selection[J]. Journal of Software, 2010, 21(2): 388-400.

[12] DAI G, WANG Y. Trust-aware component service selection algorithm in service composition[A]. 2009 International Conference on Frontier of Computer Science and Technology[C]. 2009.613-618.

[13] **徐萌**. **基于服务关系的服务组合相关技术研究**[D].**北京：北京邮电大学**, 2007.
XU M. Research on Service Relationship based Web Services Composition and Related Technology[D]. Beijing: Beijing University of Posts and Telecommunications, 2007.

[14] **叶世阳**, **魏峻**, **李磊等**. **支持服务关联的组合服务选择方法研究**[J]. **计算机学报**. 2008, 31(8): 1383-1397.
YE S Y, WEI J, LI L, *et al*. Service correlation aware service selection for composite service[J]. Chinese Journal of Computers, 2008, 31(8): 1383-1397.

[15] **李素粉**. SOBE **环境下支持信任和服务关联的服务选择方法**[D]. **北京**: **清华大学**, 2011.
LI S F. A Trust and Service Relation Aware Approach to Services Selection under SOBE Environments[D]. Beijing: Tsinghua University, 2010.

[16] [EB/OL]. http://aws.amazon.com.

[17] **张跃**, **邹寿平**, **宿分**. **模糊数学方法及其应用**[M]. **北京**: **煤炭工业出版社**, 1992.
ZHANG Y, ZOU S P, SU F. Fuzzy Mathematics and its Application[M]. Beijing: Ching Coal Industry Publishing House, 1992.

[18] **姜峰**.**基于关系的服务资源管理关键技术研究**[D]. **北京**: **清华大学**, 2009.
JIANG F. Research on Key Technologies of Relation-Based Service Resource Management[D]. Beijing: Tsinghua University, 2009.

[19] GAREY M, JOHNSON D. Computers and Intractability: A Guide to the Theory of NP-Completeness[M]. New York: W.H.Freeman and Company, 1979.

作者简介：

张云勇（1976-），**男，江苏盐城人，博士后，中国联通研究院云计算总体组主任，主要研究方向为下一代开放网络、固定移动融合核心网、移动互联网及业务和公共运算。**

李素粉（1979-），**女，河北永年人，博士，中国联通研究院工程师，主要研究方向为云计算相关技术。**

吴俊（1971-），**男，湖南长沙人，博士，北京邮电大学副教授，主要研究方向为移动互联网应用、云计算商业价值评价等。**

房秉毅（1980-），**男，山东青岛人，博士，中国联通研究院高级工程师，主要研究方向为云计算、核心网新技术。**