QoS保证的资源竞争与用户需求策略研究*)

罗 军[1] 袁 满[1] 胡建平[1] 阚志刚[2] 马 健[2]

(北京航空航天大学计算机科学与工程系601教研室 北京 100083)[1]

(诺基亚中国研发中心 北京 100013)[2]

Resource Competition and User Requirement Policy Study for QoS Guarantee

LUO Jun[1] YUAN Man[1] HU Jian-Ping[1] KAN Zhi-Gang[2] MA Jian[2]

(Beijing University of Aeronautics and Astronautics, Dept. Computer Science and Engineering, Beijing 100083, China)[1]

(Nokia R&D China, Beijing 100013, China)[2]

Abstract QoS issues are widely being studied for Internet. It is a key issue to how effectively control open resources of network in term of user's QoS need. In this paper, we formulize the resources of network and analyze its status. Further, we analyze competition for the resource of network between applications and user's need, and then point out the object of resources allocation. This paper will be useful reference for QoS study.

Keywords Internet, QoS, Network resource allocation, Resource competing, Resource compete isolation

1. 引言

服务质量(QoS)是当前网络技术研究的一个广泛的课题。服务质量是用户需求空间和网络能力空间的映射,而网络能力空间主要由网络可资源控制能力决定[1]。如何根据用户的质量需求,有效地控制开放的共享网络资源空间,是QoS保证的关键问题之一。本文通过对网络资源的形式化定义、状态分析、竞争分析以及对用户需求的形式化分析,提出了资源分配的目标、网络资源与用户服务质量等级的非一一对应关系,并分析了目前常用的竞争隔离方法一资源预留和优先级保证对网络资源利用的优缺点,讨论了当前QoS解决的基本问题,为网络的服务质量研究提供一个一般性的参考。

2. 网络系统的服务质量

一个分布式应用的系统的目标是完成用户特定任务需求。影响应用系统质量的因素主要存在两个方面,一是网络的传输质量;二是应用系统自身的控制质量能力。它覆盖端系统及端到端的网络系统。因此,应用系统的质量保证不仅覆盖应用系统自身(端系统),而且覆盖端到端的网络传输的端到端的质量保证。

网络服务质量可以基于两个方面来评价:

- 对具有不同应用特征的应用类型传输质量保证能力。
- 对于个体用户需求特征的传输质量保证能力。

网络的传输质量是影响应用质量的重要因素。但是对于不同类型的应用,其影响程度也不同,这主要由应用的具体特征所决定。具有不同特征的应用,它们对网络传输质量各方面的敏感程度也不同,因此,评价一个网络系统质量的好坏,不同的应用类型,所评判的标准有所不同。

进一步地,对于不同的用户个体,对网络传输质量的要求有所不同,所提出的相对评判标准也有所不同。从应用(用户)的角度,评价网络传输质量的好坏,是看它是否能够满足一个具体应用所要求的传输质量要求。

我们把网络对于不同质量评判标准的用户需求的满足能力,称为网络的服务质量(QoS)。网络服务质量保证的目标是能够同时满足不同的应用类型、不同用户需求的各种应用的数据传输要求。

网络的物理资源提供能力决定了网络的性能。在一个开放的共享的网络环境下,网络为应用所提供的资源处在一个不断变化的动态环境中。因此,为不同的应用提供一个稳定而有效的资源保证是网络传输质量保证要解决的根本问题。

影响应用质量的根本是网络的传输资源保证。因此,服务质量保证问题,本质上是在一个开放的分布式计算环境下基于用户需求对网络资源竞争的控制问题。传统的尽力型网络处在一个无序的、不可控的共享资源竞争状态,因此,狭义的服务质量保证要解决的根本问题是共享资源的竞争管理、控制问题。

一般情况下,主要通过两种途径:

- 控制新发起的应用对共享资源的竞争(例如:通过准入控制机制)。
- 增加网络资源的可控性,实施对网络资源的细粒度管理,使已发起的应用对共享资源的竞争处在一个有序的、可控的状态(通过资源分配、资源预留、资源调度等机制)。

从端到端的服务质量保证角度讲,共享资源包括端系统资源(计算资源、存储资源等)和网络资源(带宽资源等)。端系统上的多应用、多任务引起端系统共享资源的竞争;网络的多用户、多应用引起网络资源的竞争。当前的服务质量研究更多地集中在解决网络资源的竞争问题。

*)863项目编号:2001AA112052。此课题受诺基亚中国研发中心移动IPv6服务质量项目资助。罗 军 博士生,高级工程师,主要研究领域为分布式系统、多媒体QoS及无线通信。袁 满 博士生,副教授,主要研究领域为网络QoS,Internet服务管理、网络管理,移动IP及信息系统。阚志刚 博士,讲师,主要研究领域为移动IP,IPv6技术,多媒体QoS及网络管理等。胡建平 教授,博士生导师,主要研究领域为分布式系统、移动计算等。马 健 教授,诺基亚中国研发中心经理,主要研究领域为移动计算,无线通信及QoS管理,IPv6等。

3. 网络资源竞争与分配

用户应用的资源需求主要是端系统的计算资源、存储资源和网络的存储资源及网络的带宽资源。

在一个多应用的分布式系统环境中，存在着多应用对共享资源的竞争问题，这主要表现在大量的不同端系统对网络资源的竞争和端系统应用间对端系统资源的竞争。这种竞争势必引起任务间的相互干扰，影响应用的 QoS 保证。因此，QoS 保证主要解决的问题是多任务流对共享资源的竞争问题。这主要通过两个方面来实施资源保证：

· 竞争隔离。它主要是隔离任务流之间的相互影响，把一个任务流所面对的共享的资源环境转变为一个相对独占的资源环境。

· 资源补偿。资源的竞争将导致一个特定应用任务的资源缺乏，因此，必须对这个应用的资源进行补偿，以保证应用的基本资源需求。

相对于端系统物理资源，网络共享物理资源是比较匮乏的。因此，下面我们主要针对网络共享物理资源的状态、竞争与分配情况加以讨论。

3.1 共享资源竞争状态

设 i 表示一个分布式应用的标识；网络的物理资源为 R_N；在 t 时刻网络的每一个应用的资源需求为 $R_{A(i)}(t)$；网络为每一个应用提供的资源为 $R_N^i(t)$，理想的情况下则有 $\sum_{i=1}^{n(t)} R_N^i(t)=R_N$；在 t 时刻网络存在 n(t)个应用。网络的资源状态存在下面两种情况：

①当 $\sum_{i=1}^{n(t)} R_{A(i)}(t) \leqslant R_N$ 时，非资源竞争状态。

②当 $\sum_{i=1}^{n(t)} R_{A(i)}(t) > R_N$ 时，资源竞争状态。

状态转移可表示为：在一个确定的时间段内，资源的竞争状态和非竞争状态可能是交替出现的(图 1)，这主要是因为：

(1) 一个确定的应用的资源需求在每一时刻是不同的，即 $R_{A(i)}(t)$是时变的。

(2) 不同的时刻竞争资源的应用数量是不同的，即 n(t)是时变的。

这两个时变因素 $R_{A(i)}(t)$和 n(t)使得网络的物理资源在不同的时刻处于不同的状态，使得网络为特定应用提供的资源也在发生变化。

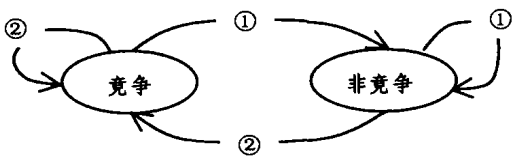

图 1

3.2 应用流间的资源的竞争与分配

在多应用的分布式环境中，处在竞争状态的共享资源主要存在两个方面的竞争：

· 已发起的应用流之间的资源竞争。这主要是应用流的资源需求 $R_{A(i)}(t)$是时变的。应用流之间的恶性竞争可能导致网络拥塞，使共享资源的利用率下降。

· 由于新应用的加入和已发起应用的退出，使得总的竞争共享资源的应用流的数量是时变的。新的应用流加入竞争，需要消耗一定的共享资源；已发起应用的退出，将释放一定的共享资源。

共享的网络资源分配的目标是：

· 保证每一个应用的资源需求。

· 保证最大化地利用共享资源。

在 t 时刻，理想的共享资源分配的目标是：

$$\begin{cases} \sum_{i=1}^{n(t)} R_n^i(t) = \sum_{i=1}^{n(t)} R_{A(i)}(t) & (1) \\ R_n^i(t) = R_{A(i)}(t) \mid i=\{1,n\} & (2) \end{cases}$$

其中：(1)式保证了最大化地利用共享资源。(2)式保证了每一个应用的资源需求。为了在每一时刻保证每一个应用流的资源需求 $R_{A(i)}(t)$，网络资源分配机制必须时刻响应每一个应用流的资源需求变化，使得 $R_n^i(t)=R_{A(i)}(t) \mid i=\{1,n\}$。

那么，在这样一个分布式环境下，能否实现这样一个精确的控制来保证这种理想的分配目标呢？从控制论的角度，我们可以把实际的共享资源分配问题进一步抽象为以 $R_n^i(t)=R_{A(i)}(t) \mid i=\{1,n\}$为目标的系统控制问题。

假设 $\vec{X}(t)$为系统的资源状态空间，

$\vec{X}(t)=\{\vec{R}_N(t), \prod_{i=1}^{n(t)} \vec{R}_N^i(t), \prod_{i=1}^{n(t)} \vec{R}_{A(i)}(t)\}$，$\vec{W}(t)$为系统的干扰，$\vec{U}(t)$为系统的控制输入，$\vec{Y}(t)$为系统输出，则系统的状态方程为：

$$\left.\begin{aligned} \vec{X}(t) &= F\{\vec{X}(t), \vec{U}(t), \vec{W}(t)\} \\ \vec{Y}(t) &= G(\vec{X}(t), \vec{U}(t)) \end{aligned}\right\} \quad (3)$$

而 $\vec{X}(t)$和 $\vec{W}(t)$不能被精确观测，所以对资源分配不能进行准确的控制。

由于实际时变因素的影响($R_{A(i)}(t)$、n(t))，以及各应用之间的相互干扰，使得实际的共享资源提供为：

$$\left.\begin{aligned} & R_N^i(t) = R_{A(i)}(t) + \Delta R_N^i(t) \mid i=\{1,n\} \\ & \sum_{i=1}^{n(t)} R_N^i(t) \neq \sum_{i=1}^{n(t)} R_{A(i)}(t) \end{aligned}\right\} \quad (4)$$

$\Delta R_N^i(t)$为网络提供给应用 i 的资源偏差。

· 当 $\Delta R_N^i(t)>0$ 时，有 $R_N^i(t)>R_{A(i)}(t)$；存在共享资源浪费。

· 设应用流 i 在 t 时刻的最小资源需求为 $R_{A(i)}^{min}(t)$，最大资源需求为 $R_{A(i)}^{max}(t)$，资源需求的波动范围为：

$$\Delta R_{A(i)}^{min}(t) = R_{A(i)}^{max}(t) - R_{A(i)}^{min}(t) \quad (5)$$

当 $\Delta R_N^i(t)<0$，且 $R_N^i(t)<R_{A(i)}^{min}(t)$时，应用的最小资源需求将不能保证。

所以，实际的资源分配目标是：

$$|\Delta R_N^i(t)| \to 0 \quad (6)$$

3.3 应用流间的资源竞争的隔离

由于应用流间的资源竞争引起了这种资源分配偏差 $\Delta R_N^i(t)$，而在这种分布式环境中不可能实施一个准确的控制来保证$|\Delta R_N^i(t)| \to 0$。所以，我们主要通过对应用流间进行资源利用的隔离，就能消除这种竞争，从而减小或消除 $\Delta R_N^i(t)$。

· *资源预留* 资源预留方法是为特定的应用流在网络上预留它所需要的资源数量，通过这种形式使这个流与其它流之间进行隔离，消除了相互干扰，形成了共享资源的部分独占。

设 $R_N^{Res(i)}$ 是网络机制为发起的应用流 i 预留的网络共享资源。应用流 i 对于这个预留资源 $R_N^{Res(i)}$ 具有独占方式，即 n(t)=1。这时对预留资源 $R_N^{Res(i)}$ 的利用率受应用流 i 在$[t_0,t_e]$

时间周期的资源需求波动的影响。当选择 $R_N^{Res(i)}=\max_{t=t_0}^{t_E}\{R_{A(i)}(t)\}$ 进行资源预留时，即按应用流 i 的最大的资源波动需求来预留资源，则总有 $R_{A(i)}(t)\leqslant R_N^{Res(i)}$，即 $R_N^{Res(i)}$ 将总是处于非竞争状态。因此，消除了应用流 i 的资源竞争。

但由于 $R_{A(i)}(t)\leqslant R_N^{Res(i)}$，在 t 时刻应用流 i 存预留资源 $R_N^{Res(i)}$ 的闲置部分为 $\Delta R_N^{Res(i)}(t)=R_N^{Res(i)}-R_{A(i)}(t)$，存在资源浪费问题。

一个极端的情况是，当网络资源 R_N 被完全预留给 m 个应用流时，即：$\sum_{i=1}^{m}R_N^{Res(i)}=R_N$，这时新的应用流将无法进入。在 t 时刻总的资源闲置为 $\sum_{i=1}^{m}\Delta R_N^{Res(i)}(t)$。

这种方法的缺点是：

· 难以准确估计每一个应用流的资源需求量，因而难于确定预留的资源 $R_N^{Res(i)}$。过多的预留资源将引起较大资源的浪费。

· 减小了网络资源的统计复用率，因而减小了网络的资源利用率。

· 要求流所经过路径上的每一个网络节点必须进行资源预留，保持每一个流的状态，大量的流状态信息间接地浪费了网络节点的存储资源和计算资源，同时也较难实现。比较典型的方案是 IETF 的 InteServ。

实施资源预留的应用流，将运行在一个相对稳定的资源环境下。

· 优先级保证　由于资源预留方法存在一定的缺点，而这种缺点主要是由于实施对每一个应用流的资源竞争逻辑上的隔离，实行资源的独占，它不能体现应用流间相互的重要程度。

优先级保证的方法，考虑了应用流间存在着的重要程度的差异，这种差异体现为不同的优先级，引入了高优先级流对低优先级流的抢占机制，提高了共享资源的利用率。

网络提供一定资源保证的优先级分类。对于每一个应用流，网络给予一定的优先级分类。网络节点通过应用流的不同优先级给予不同等级的资源保证。这样就间接地屏蔽了较低的优先级对它的干扰，相当于隔离了低优先级应用流对它的影响。

设网络提供的优先级分类 k(j)，j={1，…，p}，一个具有优先级保证 k(j)的应用流 i 表示为 $A_i^{k(j)}$，在 t 时刻它的资源需求为 $R_{A(i)}^{k(j)}(t)$，网络为具有优先级 k(j)的应用提供的资源保证为 $R_N^{k(j)}(t)$。则对于 m 个具有优先级 k(j)的应用流，在 t 时刻总的资源需求为 $\sum_{i=1}^{m}R_{A(i)}^{k(j)}(t)$。全部应用的资源需求为 $\sum_{j=1}^{p}\sum_{i=1}^{m}R_{A(i)}^{k(j)}(t)$。网络机制应当总是保证

$$\sum_{i=1}^{m}R_{A(i)}^{k(j)}(t)\leqslant R_N^{k(j)}(t) \tag{7}$$

当 $\sum_{i=1}^{m}R_{A(i)}^{k(j)}(t)>R_N^{k(j)}(t)$ 时，将抢占最低优先级应用占用的资源。当 $\sum_{j=1}^{p}\sum_{i=1}^{m}R_{A(i)}^{k(j)}(t)>R_N$ 时，最低优先级应用资源将不能被保证。

假设优先级标识号 1 为最高优先级，p 为最低优先级，一个一般的原则是高优先级的应用流首先抢占 k(p)应用流占用的资源，然后抢占 k(p−1)应用流占用的资源。

具有优先级调度的资源保证有以下优点：

· 间接地实现不同优先级类应用流之间的竞争隔离，保证了高优先级应用类的资源需求。我们可以认为高优先级类的应用流相当于实现资源独占。

· 它根据应用流的优先级把具有相同优先级的应用流进行聚类，在同一等级的资源空间内，可以进行资源的统计复用，克服了资源预留（资源独占）方式下由于每一个流的资源需求的时变性（$R_{A(i)}(t)$）而引起的资源浪费，提高了资源的利用率。

但是，由于存在着对应用流的聚类，这种聚类的限制使得应用流的 QoS 保证行为难以得到很细粒度的刻划。对于相同优先级的应用流，必须保证共享资源分配的公平性。在端系统上，如果存在多任务环境，同样存在各任务流对端系统资源的竞争问题。可以在操作系统级采用上述同样的方法来解决处理。

3.4　应用流内的应用行为特征间的资源竞争

一个用户应用的 QoS 需求主要是由应用的任务流所体现的行为特征来表现的，这种行为特征的表现转换为对端系统和网络的资源需求 $R_{A(i)}(t)$。类似于多任务流间的资源竞争，在系统仅能提供的一定的资源保证的情况下，一个应用流内的应用表现特征间也存在对共享资源的竞争问题，因此，我们也必须通过一定的方式进行这种竞争的隔离。

用户对任务流的特征表现有不同的期望值，例如：对于可视电话应用，用户对语音的保真度的期望值要高于对视频保真度的期望值，即用户希望宁愿丢失一定的乃至全部的视频保真度，也不愿意丢弃部分语音保真度。类似于多任务流间的资源竞争策略，我们也可以把用户所关心的任务流表现特征分为不同的优先级，对高优先级的行为特征进行优先的资源保证。即通过对低优先级的特征表现的降级或舍弃来释放一定的资源，从而保证高优先级的特征表现的资源需求。

设应用流 i 具有 m 个表现特征，被分为 p 类优先级，每个优先级表示为 k(j)，j∈(1，…，p)，它所对应的特征的个数为 $n^k(j)$，其中 j∈{1，…，p}，$\sum_{j=1}^{p}n^k(j)=m$。应用流特征标识为 l，每一个特征可以被表示为 $C_i^{K(j)}(l)$，其中 $l\in\{1,\cdots,n^k(j)\}$。

对于应用流 i 的每一类优先级 k(j)，它的应用特征向量

$$\begin{cases}\vec{C}_i^{k(1)}=\{C_i^{k(1)}(1),\cdots,C_i^{k(1)}(n^k(1))\}\\ \vec{C}^{k(2)}{}_i=\{c_i^{k(2)}(2),\cdots,C_i^{k(1)}(n^k(2))\}\\ \cdots\cdots\\ \vec{C}_i^{k(p)}=\{C_i^{k(1)}(p),\cdots,C_i^{k(1)}(n^k(p))\}\end{cases} \tag{8}$$

则应用流 i 的应用特征向量为：

$\vec{C(i)}=\{\vec{C}_i^{k(1)},\vec{C}_i^{k(2)},\cdots,\vec{C}_i^{k(p)}\}^T$。

设每一个应用特征 $C_i^{K(j)}(1)$ 在 t 时刻对应的资源需求为 $R_{i,1}^{K(j)}(t)$，其中 $1\in\{1,\cdots,n^k(j)\}$，j∈{1，…，p}，则为了保证应用流的全部特征 $\vec{C(i)}$ 所需要的资源为

$$\sum_{j=1}^{p}\sum_{l=1}^{n^k(j)}R_{i,j}^{K(j)}(t)=R_{A(i)}(t) \tag{9}$$

对于网络 t 在时刻为应用流 i 提供的资源 $R_N^i(t)$，当 $\sum_{j=1}^{p}\sum_{l=1}^{n^k(j)}R_{i,j}^{K(j)}(t)>R_N^i(t)$ 时，存在各应用特征 $C_i^{K(j)}(1)$ 间需求资源的竞争。假设优先级标识号 1 为最高优先级，p 为最低优先级。一般的原则是高优先级的应用特征首先抢占 k(p)特征占用的资源，然后抢占 k(p−1)特征占用的资源。

多应用流间的资源竞争隔离保证了每一个应用流的资源

需求，而一个应用流内的应用特征间的资源隔离保证了每一个应用特征的资源需求。

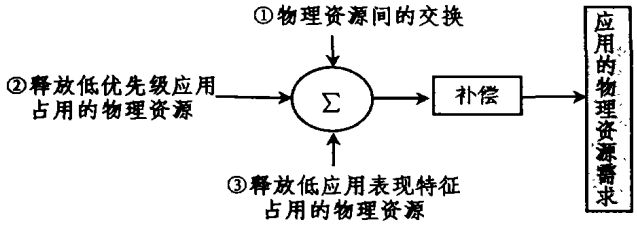

图 2　网络资源补偿

3.5　资源的释放-补偿问题

在动态的资源环境中，资源的变化将直接影响到应用的服务质量。对一个确定的应用，当它所需要的物理资源不能满足要求时，必须通过一定的方式进行资源的补偿(图 2)。

· *物理资源间的相互转换*　一个具体的应用存在对多种不同类型的物理资源的需求。同时，对一种资源需求的不足可能通过对其它资源的消耗来弥补。例如：在网络带宽匮乏的情况下，可以通过数据压缩的方式来减少带宽资源的需求，相关的代价是消耗了两个端系统的计算资源。一般情况下认为，端系统具有较为丰富的资源，而网络资源的利用则相对饱和，这是因为网络设计的原理要求最大化地利用网络资源。这种方案一般用于端系统资源和网络资源的相互转换。

· *降级与丢弃*　这种方法的本质是，高优先级应用抢占低优先级应用的资源，高优先级应用表现特征抢占低优先级应用表现特征的占用资源。对于多个应用竞争资源的情况，为了保证高优先级应用的资源需求，通过对低优先级应用的质量降级，释放一定的资源来补偿高优先级的资源需求。同样，对于一个应用中的不同表现特征竞争资源，也可以通过降低低优先级表现特征的质量需求来保证高优先级特征的质量需求。一个极端的情况是丢弃低优先级的应用和表现特征。

· *准入控制*　准入控制是根据现有的资源状态来确定一个新发起的应用是否允许被连接使用。一般的原则是首先要保持现有的连接应用，而不考虑应用的优先级高低。即在准入判别时，尽管一个新应用请求的优先级很高，但它也不能抢占正在运行的较低优先级应用的资源。

当一个应用所要求的物理资源匮乏时，可以通过图 2 所示方式进行补偿。

4. 用户的 QoS 需求

用户的 QoS 需求主要是为用户任务所实施的应用提供一定的应用表现特征的质量要求。我们把应用的表现特征分为两类：有保证的、可降级的。

· 有保证的表现特征。必须有足够的系统资源来保证应用的这些表现特征的实现。

· 可降级的表现特征。系统的资源匮乏时，这种表现特征可以根据系统资源状态进行特征性能的降级处理，直至丢弃，为有保证的表现特征提供足够的资源空间。

当一个系统的表现特征全部表现为有保证的表现特征的情况下，应用的支持系统必须能够提供足够的系统资源(端系统和网络)来保证这些有保证的应用表现特征的实现。

当一个系统的表现特征部分地表现为可降级的表现特征的情况下，可以通过对应用的表现特征的降级处理来适应网络资源状态，保证用户的服务质量需求。

一般情况下，用户的 QoS 需求是一个可接受的范围，它由有保证的特征需求和可降级的特征需求两部分组成。有保证的特征需求是基本需求，可降级的特征需求是扩展需求，是应用系统的可调整部分。在这种情况下，可以通过对低优先级表现特征的降级或丢弃来为高优先级的表现特征提供足够的系统资源。

4.1　任务的 QoS 问题形式化

我们定义用户所要完成的一件事务为一个用户任务 T。一个用户任务由 m 个任务的应用流 A_i 组成。

对于 A_i，我们给定：

$C_i=\{c_{i1},c_{i2},\cdots,c_{in(i)}\}$；$Q_i=\{\vec{q}_{i1},\vec{q}_{i2},\cdots,\vec{q}_{in(i)}\}$；

$c_{ij}\mapsto\vec{q}_{ij}\mid j=1,\cdots,n(i),i=1,\cdots,m$

C_i 是应用流 A_i 的应用特征；Q_i 是应用流的质量需求。每一个 c_{ij} 对应的质量需求为 $\vec{q}_{ij}$，$n(i)$ 为应用流 A_i 的特征个数。

对于任务 T，有：任务的应用特征向量 $C_T=C_1\times C_2\times\cdots\times C_m$，任务的质量需求向量 $Q_T=Q_1\times Q_2\times\cdots\times Q_m$，假设存在 k 类共享资源 $R_1,R_2,\cdots,R_k$，分配给任务 T 的资源为一个向量 $R_T^{Res}\in R_1\times R_2\times\cdots\times R_k$，则任务 T 的资源需求也是一个向量，$R_T\in R_1\times R_2\times\cdots\times R_k$。

进一步地，假设每一个应用特征 c_{ij} 的资源需求为 $\vec{r}_{ij}\in R_1\times R_2\times\cdots\times R_k$，则每一个应用流 A_i 的资源需求为 $\sum_{j=1}^{n(i)}\vec{r}_{ij}$，任务 T 的总资源需求为 $R_T=\sum_{i=1}^{m}\sum_{j=1}^{n(i)}\vec{r}_{ij}$。

任务 T 的 QoS 保证需要解决以下 3 个方面的问题：

(1)$F:Q_T\to R_T$。即用户提出的应用特征的质量需求如何映射到资源需求的问题。

(2)为了保证任务 T 的所有应用特征的质量需求，必须保证 $R_T^{Res}\geqslant R_T$。即资源的分配与保证问题。

(3)当 $R_T^{Res}<R_T$ 时，应用特征 c_{ij} 间的资源竞争问题。

4.2　任务的质量 Q_T 与资源 R_T 的关系

一般情况下我们总是试图找到资源与质量之间的这样一个函数，保证(6)式资源分配的目标。事实上，在多资源类型和多质量特征的情况下，这样一个函数是不存在的。这是因为应用本身可以使用两个或多个算法来实现同样的服务质量，而不同的算法对不同类型的资源需求是不一样的。对于两个具有不同压缩算法的应用 A1 和 A2，假设 A1 具有相对低的压缩比，使用较少的计算资源，而 A2 具有相对高的压缩速率，使用更多的计算资源。对于保证相同的应用质量，结果是 A1 使用了较少的 CPU 资源，使用了更多的网络带宽；A2 使用了较多的 CPU 资源，使用了较少的网络带宽。一个形式化的表述如下所示：

$$\vec{q}=\langle q_1,\cdots,q_2\rangle\begin{cases}A_1\ \vec{r}_{A_1}=\langle r_1,\cdots,r_m\rangle\\ A_2\ \vec{r}_{A_2}=\langle r'_1,\cdots,r'_m\rangle\end{cases}$$

因此，我们不能把这种情况描述为质量空间与资源空间的函数关系。

同样，给定一定的资源状态，产生不同的质量结果。

$$\vec{r}=\langle r_1,\cdots,r_2\rangle\begin{cases}A_1\ \vec{q}_{A_1}=\langle q_1,\cdots,q_m\rangle\\ A_2\ \vec{q}_{A_2}=\langle q'_1,\cdots,q'_m\rangle\end{cases}$$

因此我们也不能找到一个从资源空间到质量空间的一个映射函数。资源空间与质量空间的对应关系是一些散列点。因此，它们之间存在一定的关系，但不能用一个确定的函数来定

义。我们把它描述为一般的数学关系：$r \mapsto_i q | r \in R, q \in Q_i$。

4.3 应用特征 c_{ij} 间的资源竞争

应用特征是应用给用户提供质量的最终表现形式，用户的 QoS 需求最终表现在应用特征上。从用户的角度来看，每一个用户所关心的应用特征有所不同，这样，我们可以对不同的应用特征赋予一定的资源保证优先级。当出现资源匮乏时（$R_T^{new} < R_T$），可以通过对低优先级特征的降级或丢弃来释放一定的资源，保证高优先级特征的资源需求。

通过应用所具有的特征空间 C_T 来刻画任务 T 的服务 QoS 特征。同时任务对应用特征的资源保证具有 k 类优先级，每一个应用特征 c_{ij} 所具有的优先级为 $p(c_{ij}) = \{p_1, \cdots, p_k\}$。引入优先级标识后，我们把每一个应用特征的标识 c_{ij} 定义为 $c_{ij}^{p(c_{ij})}$，则有：

$$C_T = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_m \end{bmatrix} = \begin{bmatrix} c_{11}^{p(c_{11})} & c_{12}^{p(c_{12})} & \cdots & c_{1n(1)}^{p(c_{1n(1)})} \\ c_{21}^{p(c_{21})} & c_{22}^{p(c_{22})} & \cdots & c_{2n(1)}^{p(c_{2n(1)})} \\ \vdots & & \cdots & \\ c_{m1}^{p(c_{m1})} & c_{m2}^{p(c_{m2})} & \cdots & c_{mn(m)}^{p(c_{mn(m)})} \end{bmatrix}$$

设 $p_a | a \in \{1, \cdots, k\}$ 类优先级的应用特征集合为 $C_{P_a} = \{c_{ij}^{p(c_{ij})} | P(c_{ij}) = p_a; i = 1, \cdots, m; j = 1, \cdots, n(i)\}$，则有 $\bigcup_{i=1}^{k} C_{P_i} = C_T$。对于这些特征的资源保证的顺序是 $C_{p1}, C_{p2}, \cdots, C_{pk}$。

每个应用特征 $c_{ij}^{p(c_{ij})}$ 存在一个用户可接收的范围 $[min(c_{ij}^{p(c_{ij})}), max(c_{ij}^{p(c_{ij})})]$。我们把这个区间离散化，取 $n(c_{ij})$ 个离散点，每一个离散点的取值为 $c_{ij}(1), c_{ij}(2), \cdots, c_{ij}(n(c_{ij}))$，则这个应用特征的用户可接受的应用特征值的范围 $\Delta c_{ij} = \{c_{ij}(k) | k = 1, \cdots, n(c_{ij})\}$。

这些离散点组成了应用特征的取值空间 $\Delta C_T = \bigcup_{i=1}^{m} \bigcup_{j=1}^{n(i)} \Delta c_{ij}$。

5. 用户的 QoS 等级范围

由以上定义可知，用户的 QoS 等级的取值空间为 $Q_T^{Level} = \prod_{i=1}^{m} \prod_{j=1}^{n(i)} \Delta c_{ij}$。

5.1 用户的 QoS 等级范围

一般情况下用户选取的 QoS 等级 $Q_T^{User} = \{q_1, q_2, \cdots, q_{n(q)}\} \subset Q_T^{Level}$，这主要受到实际应用的限制。这种限制主要来自于两个方面：

- 对于用户自身的实际需求，完全的应用特征变化的组合，对于用户来讲是没有实际意义的。例如：对于视频应用，视频窗口的连续无极缩放对于用户来讲是没有意义的。
- QoS 等级的设定主要是为了应用系统来适应资源的变化，因此，必须考虑资源波动对应用的影响。每一个 QoS 等级都对应一定的资源需求，如果 QoS 等级的粒度太细，将会引起应用的抖动。

5.2 用户的 QoS 等级与应用特征优先级 $p(c_{ij})$ 的关系

在前面的讨论中，我们认为用户对不同的应用特征 c_{ij} 有不同的优先级要求。一般情况下，这种对应用特征 c_{ij} 的不同优先级要求 $p(c_{ij})$ 应该体现到用户的 QoS 等级的优先级 $p(q_i)$ 上，即认为它们存在一定的函数关系 $F: \{p(c_{ij}) | i = 1, \cdots, m; j = 1, \cdots, n(i)\} \to p(q_i)$，实际上，这种关系很难找到，因为用户的 QoS 等级的优先级是 $p(q_i)$ 用户对任务实现的服务质量的综合反映。

一般情况下，用户在选取 QoS 等级时，必须要考虑应用特征的优先级 $p(c_{ij})$，必须对高优先级的应用特征有较高的要求，保证高优先级特征的波动范围较小，即高优先级应用特征值的范围 Δc_{ij} 一般控制在一个较小的范围内。

结论 网络分布式应用服务质量保证是如何进行网络的资源控制来满足用户的服务质量需求。因此如何根据用户需求来控制网络资源是服务质量保证的关键。

由于应用的物理资源和应用数量的时变性，应用的服务资源不断地处在竞争状态。为了隔离应用间的资源竞争，可以通过资源预留和优先级分配方式。资源预留降低了网络资源的整体利用率，而优先级分配具有粗粒度的资源保证。网络资源的竞争分为应用流间的资源竞争和应用流内行为特征间的资源竞争。用户质量需求不能够与资源需求存在一个一一对应的映射关系。用户的服务等级需求存在一定的范围。高优先级应用抢占低优先级应用资源，高优先级应用特征抢占低优先级应用特征资源。

参考文献

1 Luo Jun, Yuan Man, Hu Jianping. Broad-sense QoS model for next-generation Internet. Proc. SPIE Vol. 4911, Wireless and Mobile Communications II. 2002. 265～272

2 Aurrecoechea C, Campbell A T, Hauw L. A Survey of QoS Architectures. ACM Multimedia Sys. J. - Special Issue on QoS Architecture, May 1998

3 Hutchison D, et al. QoS Management in Distributed Systems in Network and Distributed Systems Management. M. Sloman, Ed., Addison-Wesley, 1994. 273～302

4 Hutchison D, Mauthe A, Yeadon N. Quality of service architecture: Monitoring and Control of Multimedia Communications. Electronics and Commun. Engineering J., 1997, 9(3): 100～106

5 Loyall J P, et al. Specifying and Measuring QoS in Distributed Object Systems. In: Proc. ISORC '98, Kyoto, Japan, 1998

6 Campbell A, Coulson G. A QoS Adaptive Multimedia Transport System: Design, Implementation and Experiences. Media Distrib. Syst. Engineering, 1997, 4: 48～58

7 Gecsei J. Adaptation in Distributed Multimedia Systems. IEEE Multi-media, April-June 1997

8 Welling G, Badrinath B. An Architecture for Exporting Environment Awareness to Mobile Computing Applications. IEEE Trans. on Software Engineering, 1998, 24(5): 391～400

9 IETF Network Working Group: RFC 2212 Specification of Guaranteed QoS, S. Shenker, C. Partridge. R. Guerin, IETF, 1997

10 IETF Network Working Group: RFC 1633 Integrated Services in the Internet Architecture: An Overview, R. Braden, D. Clark, S. Shenker, Eds., IETF, 1994

11 IETF Network Working Group: RFC 2475 An Architecture for Differentiated Services. In: S. Blake, et al. eds. IETF, 1998