维普资讯 http://www.cqvip.com

基于信息服务社区的网格资源发现方法*)

朱 承 刘 忠 张维明 肖卫东 阳东升
(国防科技大学管理科学与工程系 长沙410073)

摘 要 本文在具有标准资源分类的前提下提出了一种非集中式的网格资源发现方法,并利用仿真对其性能和开销进行了分析验证。该方法具有良好的资源发现性能,并且开销不大,可以控制。提出了基于DHT P2P的bootstrap网络解决入口结点的获取问题,并提出了bootstrap网络上基于流言传播的负载均衡方法。

关键词 网格,资源发现,P2P,DHT,负载均衡

Grid Resource Discovery Based on Information Service Community

ZHU Cheng LIU Zhong ZHANG Wei-Ming XIAO Wei-Dong YANG Dong-Sheng
(Department of Management Science and Engineering, National University of Defence Technology, Changsha 410073)

Abstract A decentralized grid resource discovery solution is presented under predefined resource taxonomy, in which information nodes with the same type of registered resources are organized together to form information service communities (ISC), and efficient navigation between different communities is achieved by a DHT P2P based networks called bootstrap network. Periodical topology maintenance communications are used to piggyback and disseminate popular data in bootstrap network to achieve better load balance. The performance improvement of ISC based resource discovery is evaluated by simulation under different cases, and overhead is also studied.

Keywords Grid, Resource discovery, Peer-to-peer, DHT, Load balance

1 网格的发展与网格资源发现

随着网格向更大规模的发展,以及网格集成资源的丰富,网格中的资源将会具有目前P2P网络的特性:不可靠的资源和间歇性的资源参与会占到相当的比例,资源类型也会更加多种多样。实际上,网格和P2P技术在各个方面都具有很强的互补性,两者的相互融合已成为学术界的共识[1]。文[2]指出,随着网格与P2P的融合,网格资源将具有以下特点:1)大规模、分布、无集中控制:各资源在地域上分布,有不同的利益和资源管理策略,相互之间可能存在隶属的层次关系,但更多的情况下是一种相互"对等"的关系;2)动态:资源参与的模式可能有很大的差异,既有大量相对稳定的结点,也有很多的资源频繁地加入或退出;3)异构性:资源的类型和特性差异很大。

网格的这一发展趋势对网格资源发现提出了以下需求:1)不依赖集中、全局的控制;2)支持基于资源属性的搜索,且资源的属性,如CPU负载、网络可用带宽等,可能不断随时间变化;3)可伸缩性;4)支持间歇性的资源参与方式,适应资源动态参与的特点。

针对以上问题,本文提出了基于信息服务社区的网格信息结点组织与资源发现方法,用于解决非集中式网格资源发现中规模与效率之间的矛盾。

2 研究现状及相关的工作

目前网格系统中的资源发现基本上是集中式的。如,在Globus中,每个VO都有一个聚合目录来提供所有VO中资源的信息[3]。尽管用户可以通过标准的协议直接访问底层的资源信息,但是Globus中没有定义信息结点之间的组织方式与资源发现请求的处理、转发方法。Web Service使用UDDI进行服务发现,但UDDI 2.0仍然是集中式的。在最新的UDDI 3.0中,引入了"registry interaction"的概念,支持UDDI结点间更灵活的交互,但将有关交互方式与对象的决定留给管理人员。总体而言,目前网格及相关系统支持信息结点间的交互,但未提供通用的非集中式的资源发现机制。

在P2P领域,为解决动态环境下资源发现的效率与系统规模之间的矛盾,已经有大量的研究,主要思路为:1)利用结点能力,如连接带宽、计算能力等的差异,在建立结点间拓扑时体现一定的层次关系,从而将能力较差的结点排除在网络中心之外,相当于减少了网络规模;2)将具有相同或类似兴趣的结点组织在一起,使资源发现请求的传递以及资源信息的扩散更有针对性;3)使用结构化拓扑以及基于分布式哈希表(DHT)的资源查找方法[5]。

以上P2P领域的研究思路已经被借鉴到网格中,用于设计高效、可伸缩的资源发现机制:文[6]研究了基于DHT的方法,并利用Hilbert曲线来增强DHT系统对模糊匹配的支持。但是该方法仍然难于有效支持基于属性的查找,当属性取值动态变化时,开销很大[2];文[7]将网格信息结点组织成树状层次拓扑,以避免搜索中的冗余消息。但是全局范围内的层次化拓扑在动态环境下难于维护,且不符合网格中资源参与的"对等"特性;文[8]提出了结点分组的方法,但结点随机分组,大大增加了资源信息注册与更新开销。

3 信息服务社区与基于信息服务社区的网格资源发现

3.1 网格资源发现相关模型

假设1:存在网格资源的标准分类,且每一个资源都有唯一的类型。

尽管网格规模增大,资源参与方式越来越动态,但网格结

*)Supported by the National Natural Science Foundation of China under Grant Nos. 70271004(国家自然科学基金)。朱 承 博士,主要研究领域为信息管理,P2P,网格,决策支持技术。

维普资讯 http://www.cqvip.com

点需要遵循标准的接口和协议以满足不同资源集成的需要。如同 Yahoo!中的 Web 页面的分类,完全可能出现网格资源的标准分类来规范资源的注册和查找。我们考虑这一标准分类体系中最底层、不可再分的类别。若资源同时具有多种类型,则我们将其视为多个类型唯一的不同资源。

定义1 网格信息结点:实现了网格规范中资源信息注册和访问标准接口的结点。

假设2:网格中每个资源只在一个信息结点上注册,且被视为其注册结点的本地资源。

在以上假设下,在多个结点上注册的资源信息被视为对其中某个注册信息的复制。

定义2 网格资源信息组织模型:$G = (R, T, N, E, rf, D)$,其中,R 为网格中的所有资源集合;T 为网格中所有资源的类型集合:$\forall r \in R, T(r)$表示资源 r 的类型,$t(r) \in T$;N 为网格中所有信息结点集合;E 代表信息结点间的邻接关系,反映信息结点的组织结构:$E(n)$为结点 n 的邻结点集合,$E(n) \subseteq N$;rf 为资源注册函数,$rf: R \to N$,在假设1下,rf 为满射;D 代表资源信息在非注册结点上的复制或链接:$D(r)$为资源 r 的所有感知结点,$rf(r) \notin D(r)$。资源 r 的感知结点为保存有资源 r 信息复制或链接的结点。

为方便表示,定义 $LR(n)$ 为信息结点 n 上注册的本地资源集合:$LR(n) \triangleq \{r \mid rf(r) = n\}$。$LR(n) \subseteq R, R = \bigcup_{n \in N} LR(n)$。对于任意 $n_1, n_2 \in N, n_1 \neq n_2, LR(n_1) \cap LR(n_2) = \Phi$。定义 $RR(n)$为信息结点 n 上保存的所有异地资源集合:$RR(n) \triangleq \{r \mid n \in D(r)\}$。

定义3 网格资源信息更新传播模型 $U: U = (G, update(r), P)$,其中,$update(r)$表示 G 中资源 r 的变化,P 表示该资源信息的更新在 G 中信息结点间的传播策略,资源 r 的变化会引起其注册结点上的资源信息的更新,并传播到 r 的所有感知结点。对于更新信息的产生(资源 r 的注册结点)或接收结点 n,$P(update(r), n)$表示资源更新信息转发的邻结点集合,$P(update(r), n) \subseteq E(n)$。

定义4 网格资源发现请求处理模型 $L: L = (G, request(R(t, A)), S, F)$,其中 $request(R(t, A))$为查找请求,定义了要查找的资源类型 t,以及应该满足的各种属性要求 A;S 为查找起始信息结点集合;F 表示 G 中信息结点间的查找请求转发策略。对于接收到的资源查找请求,结点 n 做以下操作:若与本地注册资源或与异地资源的复制信息匹配则成功;若满足停止条件则结束,否则转发到邻结点继续查找。请求转发的策略可能是随机的,也可能根据本地记录的有关异地资源信息的链接进行。对于接收到请求的结点 n,$F(request(R(t, A)), n)$表示请求转发的邻结点集合,$F(request(R(t, A)), n) \subseteq E(n)$。

3.2 现有方法分析

现有的非集中式网格资源发现的基本方法有泛洪(flooding)、路由转发(RT)以及 NEVRLATE 等三种。

泛洪 结点以系统中其它部分结点为邻结点,形成无结构网络,或具有某种拓扑特征的网络,如树状层次化结构或超立方结构。所有接收到资源发现请求的结点除了搜索本地注册资源信息外,还向所有邻结点转发请求(请求进入的邻结点除外),直至满足结束条件。若信息结点组织为无结构网络,则泛洪的请求转发方法将产生大量的冗余信息;若是具有某种结构的拓扑,则可以利用拓扑的结构来消除或减少冗余信息。为了减少冗余的消息,有一些方法使用随机或基于启发式规则的请求转发策略[9],我们将这些方法统一视为泛洪方法的变形。在泛洪方式下,结点间不扩散资源信息的更新,且有 $\forall n \in N, RR(n) = \Phi$。本文中的泛洪不仅仅指请求转发的方式,也指与之相关的信息结点组织与资源发现机制。

路由转发[10] 结点以系统中其它部分结点为邻结点,且知道所有异地资源(指向邻结点)的路由信息。从某一结点开始的资源查找请求将根据结点保存的资源路由信息转发到相应的邻结点,并最后被路由到注册有满足条件的资源的结点。在这种方式下,需要将结点上的资源信息更新传播到其它所有结点。路由转发的方法适用于无结构或结构化的信息结点网络。

NEVRLATE[8] 信息结点被分为若干组,每一个资源都必须在所有组中的任意一个结点上进行注册,因此资源发现的时候只要搜索其中任意一个组。若将在不同结点上注册的资源视为一个资源信息在多处的复制,则我们仍然可以按照以上模型对 NEVRLATE 进行描述。结点的邻结点包括同一组内的若干其它结点,以及其上注册资源在其它组内的注册结点。从某一个结点开始进行资源发现时,使用泛洪的方法搜索同一组内的结点;对于资源更新,则需要将其散布到所有组内对应的注册结点。

除以上几种基本形式外,还有不少混合形式。如,将泛洪与路由转发相结合,将资源信息散布到注册结点小范围内的邻结点,并首先使用泛洪或随机的方式进行查找,若找到满足条件资源的信息链接或复制结点,则改用 RT 方法。由于基于 DHT 的资源信息注册和查找方法难于在属性取值动态变化的情况下有效支持基于资源属性的查找,因此我们不将其作为基本方法进行考虑。

3.3 信息服务社区与网格资源发现

定义5 信息服务社区:对于给定的网格资源集合 R,网格资源类型集合 T,网格信息结点集合 N,以及资源注册映射 rf,一个信息服务社区 $ISC(t)$是一组注册有同类资源的信息结点的集合,对于其中任意一个结点 $n \in ISC(t)$,均有1)$\exists r \in LR(n), s.t.\ T(r) = t$;2)$\forall n_1, n_m \in ISC(t), \exists n_1, n_2 \cdots n_m, s.t.\ n_{i+1} \in E(n_i), 0 \leqslant i < m$。

同一社区中的各结点均注册有同类的资源,社区拓扑保持连通。

定义6 基于信息服务社区的资源发现:给定网格资源集合 R,网格资源类型集合 T,网格信息结点集合 N,以及资源注册映射 rf,基于信息服务社区的网格资源发现满足以下条件:1)$\forall n \in N, r \in LR(n)$,有 $n \in ISC(T(r))$;2)$\forall n \in N, t \in T, \exists n' \in E(n), s.t.\ n' \in ISC(t)$;3)$\forall r \in R, D(r) \subseteq ISC(T(r))$;4)对于接收到资源发现请求的结点 n,$F(request(R(t, A)), n) \subseteq (E(n) \cap ISC(t))$。

在基于信息服务社区(ISC)的网格资源发现机制中,结点根据注册资源类型加入相应的社区,且对于每一个社区,都至少了解其中的一个结点,我们称为社区入口结点;资源信息复制或链接局限于同社区的结点,只向与请求资源类型相同的社区内的结点转发请求。社区内的结点组织和请求转发方法可以使用各种现有的方法。如,将同一社区内的结点组织成无结构化网络,利用泛洪进行资源发现请求的转发。

与目前的方法相比,基于 ISC 的方法通过限制搜索规模以及资源信息更新的传播规模改善查找以及资源更新的性能,因为注册同一类型资源的结点被组织在一起,一旦获取了所要查找资源所在社区的入口结点,搜索将仅限于拥有相应类型资源的结点。因此,基于 ISC 资源发现性能的关键影响因素在于:1)搜索社区的大小;2)有效获取相应社区的入口结点。搜索社区的大小与具体问题有关,查找资源所在社区越小,基于 ISC 的方法的性能越好。对于社区入口结点的获取,

维普资讯 http://www.cqvip.com

我们在3.4节提出了基于DHT P2P的入口结点注册与查找网络——bootstrap网络，能够保证高效地获取指定社区的入口结点。

基于ISC的资源发现方法的性能提高是以增加结点拓扑维护开销为代价的，因为结点可能同时加入多个社区，必须维护更多的邻结点。我们将对开销进行分析。在现实中，一个信息结点上注册的资源种类与全体资源种类相比是很少的，因此大多数结点的存储空间和拓扑维护开销不会增加太大。

3.4 基于DHT P2P的bootstrap网络

在分布式的动态环境下，结点之间需要有相互发现的机制，来保证及时集成新结点，避免结点的退出或失效对系统造成的影响。这一问题在文[2]中称为结点发现(peer discovery)。

3.4.1 *P2P网络中的结点发现* 结点之间通过周期性的通信，如Gnutella中的ping / pong消息、Chord中的拓扑稳定操作，以及gossip等手段来获取其它结点的信息，并根据这些信息来建立并维护与其它结点的邻接关系。如，在Gnutella网络中结点之间通过ping/pong消息的交互来了解网络中的其它主机，并作缓存，称为本地结点缓存(local host cache)。Gnutella网络中还存在几个固定的结点，搜集、提供网络中其它结点的信息。为弥补本地结点缓存的不足，缓解集中式结点搜集与查找的性能瓶颈，Gnutella中提出了GwebCache协议[11]。其基本做法是架设少量(数百个)运行GWebCache脚本的服务器，其它结点通过Web协议与之交互，实现结点信息的注册和查找。

3.4.2 *基于DHT P2P的社区入口结点注册和查找网络——bootstrap网络* 基于ISC方法中入口结点的获取问题与P2P网络中的结点发现问题的差异主要在于，前者需要获取属于特定社区结点的地址信息。仅仅依靠结点的本地缓存不一定能找到特定类型的结点，采用集中式的注册和查找结点会遇到性能瓶颈，而在网络上进行分布式的搜索又会带来新的开销，且不一定能保证性能。我们的解决办法是将网格上的少量信息结点组织构造成信息服务社区入口结点注册和查找网络，称为bootstrap网络。

bootstrap网络中的结点使用DHT P2P协议组织，信息服务社区在bootstrap网络上的注册结点根据服务社区名通过哈希确定，注册结点保存有不断更新的社区入口结点的信息；而入口结点的查找则根据服务社区名通过DHT的查找协议进行。DHT网络的路由查找协议能够以很高的概率保证从bootstrap网络中的任意一点开始在$O(\log N)$步内找到对应的注册结点，N为bootstrap网络的规模。网格中的结点只要记录bootstrap网络上任意一个结点的信息即可保证随时查找特定社区的入口结点。

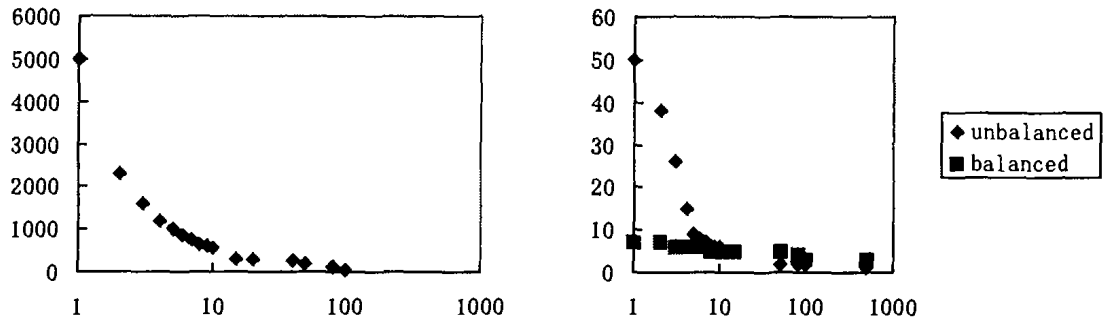
bootstrap网络可视为对GwebCache协议的扩展：利用DHT协议将原本相互之间没有连结的GwebCache结点组织起来，使各服务社区的注册和查找负载根据社区名哈希到这些结点上，并利用DHT的查找协议支持在bootstrap网络上的基于社区名的高效查找。

3.4.3 *bootstrap网络上的负载与负载均衡* bootstrap网络的负载主要是结点注册/更新以及查找负载。可以通过结点注册/更新周期的自适应的调整来控制bootstrap网络的注册/更新负载，为了控制bootstrap网络的查找负载，可以尽量利用结点本地缓存信息：当从网格中某一个信息结点开始发出查找时，首先搜索本地的结点缓存；若未找到所需服务社区内的入口结点，则小范围内通过flooding搜索邻近的结点及其缓存；若仍未找到，再向bootstrap网络提交查询。

Bootstrap网络上另一个重要的问题是负载均衡问题。对于注册/更新负载，可以使用现有的负载均衡方法[12~13]，如多哈希方法，或将负载转移到轻载结点。对于查找负载，在bootstrap网络上各结点缓存热点社区的入口结点信息有助于降低对其注册结点的查找访问数量。但是在动态环境中，可能仍然需要频繁地更新缓存。我们提出基于流言传播的方法，利用DHT P2P网络中结点周期性拓扑维护通信夹带、传播查找量大的社区入口结点信息。其基本思想为：若bootstrap网络上的结点发现其负责的某个社区入口结点的访问量很大，可以利用周期性的拓扑维护通信中向其通信对象主动推送该热点社区的最新入口结点信息；而接收到信息的结点将刷新本地缓存，并同样在周期性的拓扑维护通信中向其通信对象进行推送，从而实现热点社区入口结点信息在bootstrap网络上以后台方式进行的主动推送，且不引起额外的通信开销。

4 仿真

我们设定信息结点总数为10,000，资源总数为50,000，资源类型总数为1,000。为了评价基于ISC的资源发现方法在不同社区规模下的效果，我们考虑各类型资源数量的分布不均匀，对应的社区大小分布如图1(a)所示。与文[2]类似，我们考虑资源在各结点上分布的两种情况：均匀和非均匀，对应的结点参与社区数量的分布如图1(b)所示。



(a) 信息服务社区大小的分布　　(b) 结点加入社区数量的分布

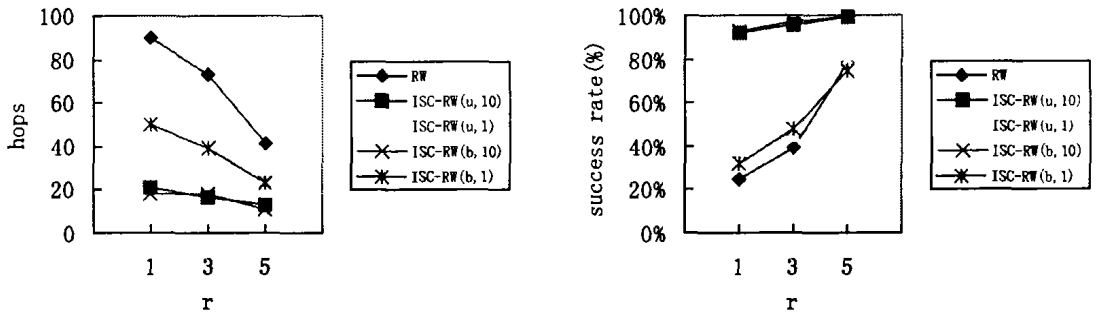
图1 网格仿真环境的设定

我们考虑ISC的方法在不同的社区规模、结点资源分布以及资源密度下的效果。这里的资源密度指拥有满足条件资源的信息结点占总信息结点的比例。为考察不同的社区规模对基于ISC方法的影响，我们分别比较当查找资源落在最大社区以及落在规模大小排名第10的社区时的情况。在我们的仿真中，根据图1，最大的社区规模约为5,000，大小排名10的社区规模约为500。

我们利用PLOD[14]算法生成所有的初始拓扑，且参数设

维普资讯 http://www.cqvip.com

为 $\alpha = 0.3$，结点度均值为4，用于模拟具有 power-law 特征的大规模网络。为控制仿真的开销，我们不对 bootstrap 网络进行实际的模拟，而是直接使用 Chord[15]系统的仿真结论：对于每个提交到 bootstrap 网络的关于入口结点的查找，我们设定成功返回的跳数为5。这是因为根据 Gnutella 网络的经验，我们期望 bootstrap 网络的规模不超过1000，而 Chord 系统中的平均查找跳数为 $1/2(\log_2 N)$，N 为 bootstrap 网络规模。我们同时根据 Chord 仿真中观测到的均值设定 bootstrap 网络上查找失败的比例为5%。当需要查找特定社区入口结点时，查找初始结点将首先查找自己的本地缓存，以及2跳内邻结点及其缓存，若仍未找到，才向 bootstrap 网络提交查找。我们使用与文[11]类似的 ping/pong 模式，但在 pong 消息中加入结点所属社区 ID 信息（最多5个）。当特定社区的入口结点无法获取时，资源发现请求将在社区间作随机的转发。资源发现请求的初始结点在不具有满足查找条件资源的结点中随机选择，在查找请求发出之前，网络有10分钟的时间进行稳定。



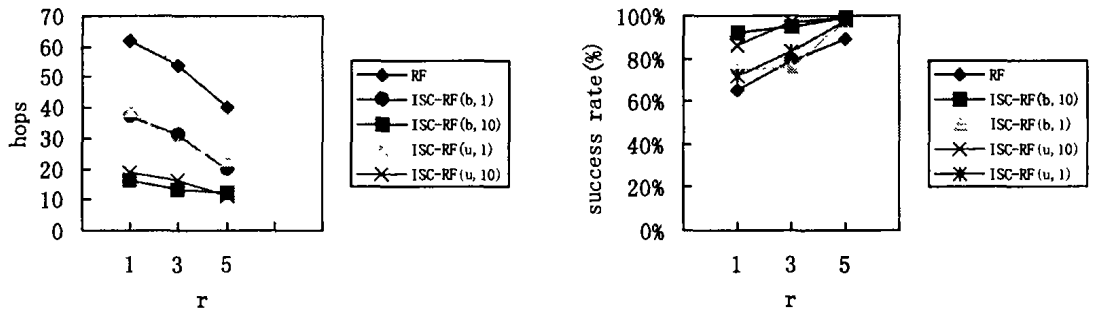
(a) 跳数　　(b) 成功率

图2　在 random walk 请求转发下，基于 ISC 的资源发现方法与泛洪的比较

4.1 性能比较

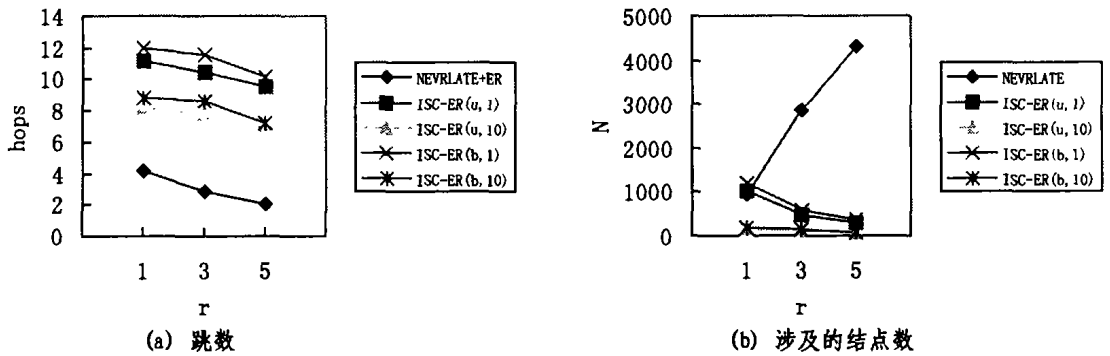
我们首先比较基于 ISC 的方法与泛洪，两者均采用 random walk[9]作为请求转发方法，结果如图2所示。其中“ISC-RW(u, 10)”代表：基于 ISC 的资源发现方法，并在社区内采用 random walk 的请求转发策略，资源在结点的分布不均匀，且所查找结点落在规模大小排名第10的社区内。其它符号的意义可以类推。横轴的“r”代表资源密度，单位为1/1000。纵坐标记录了当拥有满足资源的资源被第一次找到时的跳数。查找成功率如图2(b)所示。

类似地，我们比较基于 ISC 的方法与泛洪和 RT 的混合方法。泛洪和 RT 的混合方法指仍然使用 random walk 作为请求转发方法，直到拥有满足条件资源被找到，或其链接被找到。若是链接被找到，则沿着链接进行请求转发。在社区内同样采用混合方法，且资源信息的扩散范围均设为2跳以内的邻结点，结果如图3所示。



(a) 跳数　　(b) 成功率

图3　在 random walk 请求转发下，基于 ISC 的资源发现方法与泛洪和 RT 混合方法的比较



(a) 跳数　　(b) 涉及的结点数

图4　在 expanding ring 请求转发下，基于 ISC 的资源发现方法与 NEVRLATE 方法的比较

NEVRLATE 的性能与其分组模式有关。我们根据文[8]设定仿真中 NEVRLATE 的最佳分组方案：信息结点分为100个组，每组100个结点。我们采用 expanding ring[9]作为社区内/组内的请求转发方法，除跳数外，还比较资源发现操作

维普资讯 http://www.cqvip.com

和资源信息更新操作中涉及到的结点数。涉及到的结点越多，则方法的消息通信开销越大，结果如图4所示。

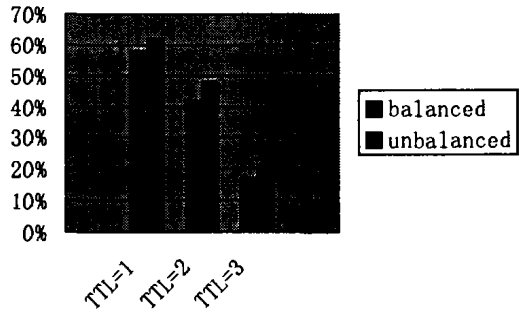
4.2 开销

基于ISC资源发现方法的主要开销来自于结点的拓扑维护，因为结点可能同时加入多个社区。而结点拓扑维护开销主要体现在结点间周期性的通信。

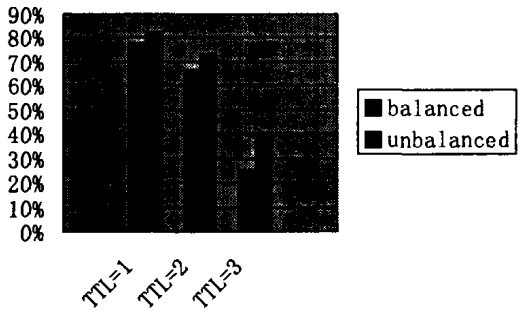
按照文[11]的方案，对于每一个连结，每3秒发送一个ping消息，并从邻结点收到pong消息以及另外10个其缓存的pong消息。每个ping消息包括：IP地址(4 bytes)、端口(2 bytes)、hops(1 byte)；而每个pong消息，则包括：IP地址(4 bytes)、端口(2 bytes)、hops(1 byte)以及最多5个所属社区ID(5 * 4 bytes)。计入包头等开销，一个ping消息约为20 bytes，而pong消息约为40 bytes，因此，每个连结上的带宽开销为：(20 + 40 * 11) / 3 ≈ 153 bytes/sec。一个社区中结点的平均连结数约为4，故加入一个社区所需的带宽开销约为612 bytes/sec。即使结点同时加入50个社区，带宽开销约为30k/sec，而且还可以通过延长ping的间隔来进一步减小开销。

4.3 bootstrap网络的负载

我们观察在一段时间内访问bootstrap网络的资源发现请求次数与总资源发现请求次数之比，及其与查找结点本地缓存搜索范围的关系，如图5所示。这表明，bootstrap网络的查找负载可以通过设定不同的本地结点缓存搜索范围进行有效控制。



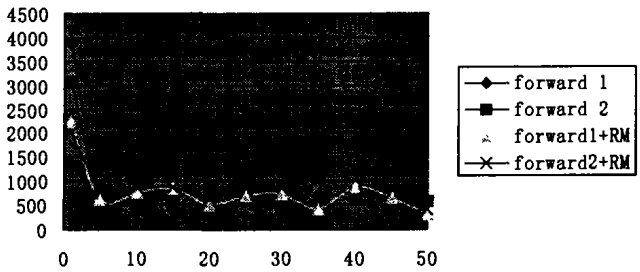
(a) 查找资源属于规模最大的社区



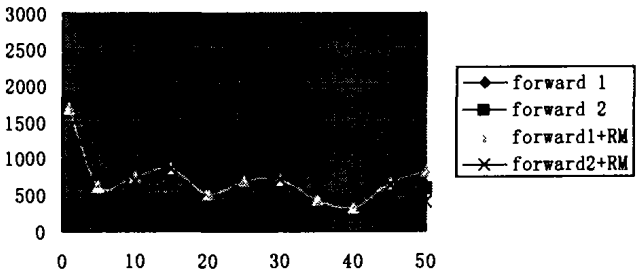
(b) 查找资源属于规模大小排名10的社区

图5 访问bootstrap网络的资源发现占总数的比例，分为结点资源分布均匀与不均匀两种情况

4.4 Bootstrap网络的负载均衡



(a) μ 为 1/300.



(b) μ 为 1/600.

图6 bootstrap网络上各结点分组的查找操作数量

我们单独仿真bootstrap网络来评价流言传播方法的负载均衡效果。设定bootstrap网络按Chord协议组织，规模为500，每个入口结点信息保持有效的时间服从负指数分布，我们设定参数分别为1/300和1/600。一段时间(30分钟)内，随机地从bootstrap网络中的其它结点出发，发出对某个特定社区入口结点总数达30000次的查找。查找出发的bootstrap结点首先检查自己的本地缓存，若有要查找的入口结点信息，且仍然有效，则将结果返回；否则按Chord查找协议转发查找请求。我们考虑两种请求的转发方式：1)查找路径的中间结点只是转发请求；2)若中间结点的缓存中有所需的入口结点信息，且有效，则查找成功返回。我们在仿真中记录每个结点的处理的查找请求次数。为便于显示，我们将bootstrap网络上在拓扑上相邻的结点每10个分为一组，第一组包括被查找结点及其后继结点，累计并显示各组处理查找请求的总数，如图6所示。其中，forward 1表示前一种请求转发策略，"+RM"表示进行流言传播时的情况。结果显示，基于流言传播的方法在结点动态变化的情况下有助于进一步降低"热点"的负载。

结论 我们在具有标准资源分类的前提下提出了一种通用、非集中式的网格资源发现方法，并利用仿真对其性能和开销进行了分析验证。结果表明，该方法具有良好的资源发现性能，且开销不大并可以控制。在该方法的实现中，我们提出了基于DHT P2P的bootstrap网络解决入口结点的获取问题，并提出了bootstrap网络上基于流言传播的负载均衡方法。

参考文献

1 Foster I, Iamnitchi A. On Death, Taxes, and the Convergence of Peer-to-Peer and Grid Computing. In: Proc. of the 2nd Intl. Workshop on Peer-to-Peer Systems (IPTPS'03). Heidelberg: Springer-Verlag, 2003

2 Iamnitchi A. Resource Discovery in Large Resource-Sharing Environments:[PHD thesis]. University of Chicago, 2003

3 Czajkowskiy K, Fitzgeraldz S, Foster I. Grid Information Services for Distributed Resource Sharing. In: Proc. of HPDC-10, IEEE press, 2001

4 Sripanidkulchai K, Maggs B, Zhang H. Efficient Content Location Using Interest-Based Locality in Peer-to-Peer Systems. In: Proc. of INFOCOM 2003, IEEE press, 2003

5 Sylvia R, Scott S, Ion S. Routing Algorithms for DHTs: Some Open Questions. In: Proc. of the 1st Intl. Workshop on Peer-to-Peer Systems (IPTPS '02), Heidelberg: Springer-Verlag, 2002

6 Andrzejak A, Xu Z. Scalable, Efficient Range Queries for Grid Information Services. In: Proc. of IEEE P2P 2002, IEEE press, 2002

维普资讯 http://www.cqvip.com

高可用的网络备份系统的研究与设计*)

谢长生[1] 韩德志[2] 李小玉[1]

(华中科技大学计算机学院 武汉430074)[1] (暨南大学计算机系 广州510632)[2]

摘 要 随着网络数字信息爆炸性增长和关键应用需求,大容量、高可用性的存储系统成为存储领域的研究热点。针对这种情况,我们将多个NAS融合成NAS簇(NASC: NAS cluster),以满足海量存储需求;为保证高可用性的需求,我们开发了一个NASC网络备份系统(简称NASCC),使用户应用关键性数据可同时镜像到NASC中的两个或更多个NAS节点,也可将NAS中备份数据恢复到用户指定的设备。并且,当某个NAS节点出现故障时,其冗余备份NAS自动接替其备份/恢复任务,保证用户备份/恢复数据的高可用性。本文将重点描述NASCC的设计和实现。

关键词 附网存储,簇,堆叠式文件系统,网络备份,高可用性

Study and Design on NAS Copy System of High Availability

XIE Chang-Sheng[1] HAN De-Zhi[2] LI Xiao-Yu[1]

(School of Computer Science, Huazhong University of Science and Technology, Wuhan 430074)[1]

(Computer Department of Jinan University, Guangzhou 510632)[2]

Abstract With the data information of volatile increase and the demand of application, greater capacity, high availability storage system has been the study hotspot of network storage. To solve the above-mentioned issues, we merge multi-NAS storage server into a big NAS cluster (NASC: NAS cluster) by storage virtualization. To guarantee high-availability and high-security for user' data, we have designed a NASC copy system, it can effectively copy user' data to NASC' NAS, or resume data from NASC' NAS. The paper emphasively discusses the design and implementation of NASC copy system.

Keywords Network attached storage, Cluster, Stackable file system, Network copy, High availability

1 引言

网络数据信息爆炸性地增长及应用数据高可用性和安全性的要求,刺激了网络备份技术的发展。目前,网络备份技术正受到很多用户的关注,不少用户已建立或已开始建立自己的网络备份系统。数据备份是存储系统最重要应用之一,也是保证存储网络高可用性的必要手段。网络备份有多种实现形式,可从不同的角度对备份进行分类,从备份模式来看,可以分为物理备份和逻辑备份。物理备份又称为"基于块(block-based)的备份"或"基于设备(device-based)的备份"。逻辑备份也可以称作"基于文件(file-based)的备份"。网络备份多与镜像技术相结合来实现。为实现海量存储,我们采用聚集技术将多个支持iSCSI协议的NAS[1,2],融合成一种NAS簇(NAS Cluster,简称NASC)[3];为保证NASC的安全性和高可用性,我们设计并实现了一种NASC网络备份系统(简称NASCC)。NASCC除通过备份/恢复保证用户数据的安全性和可用性外,还通过高可用技术来保证整个NASC中用户所备份数据的高可用性。NASCC对文件I/O请求采用的是逻辑备份,而对块I/O请求采用的是物理备份。

NASCC是基于网络环境下的高可用性数据备份/恢复系统,它有以下特点:(1)采用镜像方式,同时将用户数据写入NASC中的两个NAS中(一个称为主备份NAS,一个称为从备份NAS),这可保证用户数据的安全性;(2)可以按照用户的需要,方便、安全、完整地将数据从主备份NAS中恢复到用户指定的机器中,以方便用户的访问;(3)当主备份NAS出现故障时,从备份NAS可即时接替主备份NAS,或将数据恢复到主备份NAS中。这样可以保证用户数据的高可用性;(4)根据一定的策略,定期对NASC中的文件进行配客管理,以提高整个NASC系统的存储空间的利用率。本文将从NASCC的设计机制,以及关键设计模块入手,论述NASCC的设计与实现。

*)基金项目:本文受国有自然科学基金项目的资助,编号:60173034.谢长生 博士导师,研究方向:网络存储系统,基于IP的SAN.韩德志 博士研究生,研究方向:网络存储系统,基于IP的SAN。

7 Huang L, Wu Z, Pan Y. A Scalable and Eeffective Architecture for Grid services' Discovery. In: Proc. of WWW 2003

8 Chander A, Dawson S, Lincoln P, et al. NEVRLATE: Scalable Resource Discovery. In: Proc. of CCGrid 2002

9 Lu Q, Cao P, Cohen E. Search and Replication in Unstructured Peer-to-Peer Networks. In: Proc. of ACM SIGMETRICS 2002

10 Li W, Xu Z, Dong F, et al. A Grid Resource Discovery Model based on the Routing-Transferring Method. In: Proc. of 3rd Intl. Workshop on Grid Computing, 2002

11 http://rfc-gnutella.sourceforge.net/src/rfc-0_6-draft.html

12 Rao A, Lakshminarayanan K, Suranaet S, et al. Load Balancing in Structured P2P Systems. In: Proc. of the 2nd Intl. Workshop on Peer-to-Peer Systems (IPTPS '03), Heidelberg: Springer-Verlag, 2003

13 Byers J, Considine J, Mitzenmacher M. Simple Load Balancing in Distributed Hashing Tables. In: Proc. of the 2nd Intl. Workshop on Peer-to-Peer Systems (IPTPS '03), Heidelberg: Springer-Verlag, 2003

14 Palmer C R, Steffan J G. Generating Network Topologies that Obey Power Laws. In: Proc. of the IEEE Globecom'00, San Francisco, 2000

5 Stoica I, Morris R, Karger D, et al. Chord: A Scalable Peer-To-Peer Lookup Service for Internet Applications In: Proc. of ACM SIGCOMM 2001