计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
46(2): 186-193, 2009

数据网格中请求呈现分组特性的副本管理策略研究

姜建锦[1,2,3]　　杨广文[1,2]

[1](清华大学计算机科学与技术系　北京　100084)

[2](清华信息科学与技术国家实验室　北京　100084)

[3](北京电子科技学院计算机科学与技术系　北京　100070)

(jiangjj02@mails.tsinghua.edu.cn)

Replication Strategies in Data Grid Systems with Clustered Demands

Jiang Jianjin[1,2,3] and Yang Guangwen[1,2]

[1](*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

[2](*Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing* 100084)

[3](*Department of Computer Science and Technology, Beijing Electronic Science and Technology Institute, Beijing* 100070)

Abstract In data grid systems, data usage pattern plays an important role in system performance. According to some recent traces about real systems, data request and replica distribution exhibit clustering properties. Considered in this paper is the relationship between request distribution and replica distribution in data grid where request exhibits clustering properties. First the formal model of replication strategies in federated data grid system is given. The performance metrics include cumulative hit ratios and average access latency. Then investigated is what is the optimal way to replicate data with the objective of minimizing average access latency when request exhibits clustering properties. In the sense of minimizing average access latency, it is found that the more popular a file in a subgrid, the more replicas should be created in this subgrid; furthermore, when requests distribute uniformly in system, replicas should be uniformly distributed in system too. The optimization model is solved by means of Lagrange multiplier method and bisection method. Then, an optimization downloading replication strategy for clustering demands is obtained. The performance of this strategy is compared with that of uniform replication strategy, proportional replication strategy, square root replication strategy and LRU caching strategy through simulation. Simulation results validate the effectiveness of optimal strategy. Compared with these popular strategies, the optimal strategy has advantages of least wide area network bandwidth requirement and least average access latency.

Key words data grid; clustering; replication strategy; access latency; optimized distribution

摘　要 在数据网格中,数据使用模式将影响系统性能.根据一些实际系统的测试结果,数据请求呈现出分组特性.为研究当数据请求呈现分组特性时请求分布与副本分布的关系,首先定义了数据网格中副本复制策略的模型,然后研究在数据请求呈现分组特性时平均访问延迟最小的最优策略.采用拉格朗日乘子法以及二分法对上述模型进行求解,得到了一个在请求分组模式下的最优下载副本策略.通过模拟实验对最优策略以及均匀复制策略、比例复制策略、平方根复制策略、LRU缓存策略的性能进行了比较.结果表明,最优策略所需广域网带宽最少,平均访问延迟最小.

关键词 数据网格;分组;副本策略;访问延迟;最优分布

中图法分类号 TP393

收稿日期:2008-01-14;**修回日期**:2008-07-02

基金项目:国家自然科学基金项目(90412006, 90412011, 60573110, 90612016, 60673152);国家"九七三"重点基础研究发展计划基金项目(2004CB318000, 2003CB317007);国家"八六三"高技术研究发展计划基金项目(2006AA01A101,2006AA01A108,2006AA01A111)

数据网格的关键问题之一是数据副本管理与存储资源管理[1-2].通过合理放置数据副本,将数据传输尽量限制在物理位置邻近(或通信延迟低、带宽高)的结点之间,可以降低用户访问数据的延迟,并减少对数据网格中广域网带宽的占用,最终提高数据密集型任务的吞吐率.

在目前的数据网格副本管理研究中,大多是采用缓存策略对存储空间进行管理[3-5],即通过记录数据的访问轨迹,决定存储哪些数据的副本,以提高数据命中率,降低访问延迟.这些策略基于数据访问请求序列具有时间局部性,即当前访问的数据在随后的操作中仍然可能被再次访问.一些研究工作引入经济模型判断是否在一个结点上创建某个数据的副本[3],当一次文件请求发生时,每个结点根据自己对文件的访问请求以及对未来的预测决定在缓存中是否保留该数据.另外,一些研究工作针对文件集合进行管理[4-5],即缓存替换或准入的粒度不是一个文件,而是多个文件组成的文件集合.

上述工作中,系统中数据分布随着每次数据访问发生变化,关键问题是采用何种算法保证缓存内容在后续请求中的命中率.另外,其决策是从单个结点的局部信息出发,并不能保证全局最优.由于结点存储容量有限,不可能存储所有所需数据,总会发生数据缺失,此时就需到其他结点获取数据.而数据缺失时所造成的数据延迟大小,以及数据缺失的比例将会决定系统的整体性能.因此,一个需要解决的问题是从系统全局角度看,当数据副本服从何种分布时,访问一个数据时期望延迟最小.本文不研究如何管理结点的存储空间,而是从系统全局的数据分布出发,研究在平均访问延迟最小目标下的数据副本最优分布.

与本文相关的工作是对等网络(peer to peer,P2P)中关于数据分布的研究.一些研究工作针对对等网络中数据副本或数据索引的最优分布进行了研究[6-9],在采用随机走步(random walk)方法进行搜索的非结构化对等网络系统中,以搜索性能作为优化目标,得出的最优分布是平方根复制[6-7],即数据复制比率与数据请求比率的平方根呈正比关系.若系统的链路层网络拓扑满足幂次法则(power law),且参与到对等网络中的结点均匀分布在底层网络,以文件下载所经过的路径所对应的链路层连接数作为下载一个文件的带宽开销指标,则采用比例复制,即数据复制比率与该数据被请求的比率呈正比,可使系统的平均下载带宽开销最小.

上述工作假设针对一个文件的数据请求在系统中是均匀分布的.但是,根据近期对一些数据网格[5,10]以及 P2P 系统[11-12]进行的测试表明,对数据的请求并非均匀分布在系统中,处于不同集合中的结点对系统中文件的请求频度并不相同,本文将其称为请求分组(clustering)特性.

一些研究者从搜索角度对 P2P 系统中请求呈现分组特性的最优副本分布进行了研究.在文献[9]中,对于一个文件的请求分为高密度区与低密度区两种请求模式,并且所有低密度区的结点对于一个文件的请求模式相同.文献[9]的结果表明,最优搜索性能所对应的分布是比例复制;若查询负载最优,对应的副本分布是平方根复制.尽管该文针对一般情形进行了研究,但并未给出数据副本分布与请求的对应关系.

本文针对上述副本策略在数据网格中的性能进行了分析与模拟实验,结果表明,这些策略在广域网带宽占用、数据平均下载时间上并非最优.

本文选择联合型数据网格[13]中的数据副本分布作为研究对象,以数据访问延迟的期望作为优化目标,建立了相应的模型,并对该模型进行分析、求解.按照本文所给出的分布进行数据复制,不仅能降低对广域网带宽的占用,而且可降低数据平均访问时间,从而提高数据网格中数据处理任务的吞吐率.

1 联合型数据网格

本文针对联合型数据网格[13]进行研究.系统中共有 G 个子网格 g_i,每个子网格的结点数为 m.每个子网格中有一个超级结点,负责存储该子网格中数据副本的索引信息,并负责响应来自所属子网格其他结点的查询、以及来自其他子网格所转发而来的查询消息.这些超级结点之间采用对等方式进行互连,共同完成索引信息查询.

系统中有 M 个结点 n_k,其中 $k=1,2,\cdots,M$,这些结点分别属于不同的子网格.不失一般性,子网格 g_i 中的结点分别为 $n_{(i-1)m+1},n_{(i-1)m+2},\cdots,n_{i\times m}$,给定一个结点 n_k,它所在的子网格为 g_i,$i=\lceil k/m \rceil$.

系统共有 N 个互不相同的数据项(或文件)f_j,其中 $j=1,2,\cdots,N$.为方便讨论,假设每个文件大小相同,且每个结点存储空间相同,可存储 K 个文件.数据网格所有结点可存储 MK 个数据副本.

为讨论简便并结合实际测试结果,假设对于一个文件 f_j 而言,有两种请求模式:f_j 在一些子网格

中拥有较高的请求频率,将这些子网格组成的集合记做 GH_j,即 $GH_j=\{g_i \mid g_i$中的结点对文件 f_j有较高的请求频率$\}$,本文只讨论 GH_j包含一个子网格的情况,即$|GH_j|=1$,其他情况可类似处理,在此忽略. f_j在其他子网格中拥有较低的请求频率,将这些子网格组成的集合记做 GL_j,即 $GL_j=\{g_i \mid g_i$中的结点对文件 f_j有较低的请求频率$\}$,$|GL_j|=G-1$. GH_j每个结点平均请求频率为 λ_{hj},GL_j中结点平均请求频率为 λ_{lj},且 $\lambda_{hj}>\lambda_{lj}$. 系统中每个结点对文件 f_j的平均请求频度为 λ_j,可以得出 $M\lambda_j=m\lambda_{hj}+(G-1)m\lambda_{lj}$,每个结点对所有数据的累积访问频度为 $\lambda=\sum\lambda_j=1$. 称 $c_j=m\lambda_{hj}/(M\lambda_j)$为 f_j的聚集比例(clustering ratio),即处于 GH_j中结点发出的请求所占系统中所有结点针对 f_j的请求的比例. 此处需说明的是,由于模型中已经考虑了不同子网格间结点对数据请求的不均匀性,因此对于子网格内部假设每个结点对一个文件的请求均匀分布. 在后续的研究中,将会针对最一般的情形进行研究,即不仅子网格之间对数据请求模式不同,而且子网格内部结点的请求模式也未必相同.

数据 f_j有 r_j个副本,且每个结点最多存储 f_j的一个副本. 假设副本在子网格内部均匀分布,在高密度子网格中,拥有副本数目为 r_{hj},在低密度子网格中,拥有副本数目为 r_{lj},且 $r_{hj}>r_{lj}$,$r_j=r_{hj}+(G-1)r_{lj}$. 对于高密度子网格的一个结点而言,拥有文件 f_j的概率为 $\rho_{hj}=r_{hj}/m$,对于低密度子网格的一个结点而言,拥有文件 f_j的概率为 $\rho_{lj}=r_{lj}/m$,可见 $0\leqslant\rho_{lj}\leqslant\rho_{hj}\leqslant 1$.

数据副本的查找方式不是本文讨论的范围,假定只要系统中存在一个数据的副本,必然可以由一种策略查找到. 另外,参加到数据网格中共享的数据至少有一个副本.

假设系统中结点(包括系统规模、结点资源状况)、网络连接在某段时间内相对稳定,它们共同协作完成数据处理任务. 每个任务需要对一个或者多个文件进行处理. 若本地结点存储了这些文件的副本,任务可以执行;否则,需到系统中其他结点获取数据之后,再执行数据处理任务.

当用户访问某个文件时,若当前结点没有该文件的副本,则通过本地超级结点查看是否在本地子网格中有相应的副本,优先使用本地子网格中的数据副本;否则,通过超级结点之间的合作选择其他子网格中一个结点下载所需数据的副本. 这样,对于一次数据请求,可能在本地结点命中,或者是本地子网格其他结点命中,或者是其他子网格的结点命中.

2 副本策略相关问题

2.1 性能指标

根据结点 n_k(结点 n_k属于子网格 g_i,$i=\lceil k/m\rceil$)访问文件 f_j的过程,考虑 3 个事件,并计算各自概率:

1) n_k命中 f_j,记做 El_{kj};根据前面对副本数目及副本概率的论述,有 $P(El_{kj})=\begin{cases}\rho_{hj}, & g_i\in GH_j,\\ \rho_{lj}, & g_i\in GL_j.\end{cases}$

2) n_k未命中 f_j,但所属子网格 g_i命中,记做 Eg_{kj};

3) 子网格 g_i未命中 f_j,必定在子网格 g_i之外的子网格命中,记做 EG_{-kj};EG_{-kj}意味着子网格 g_i没有 f_j的副本,因此概率为 $P(EG_{-kj})=[1-P(El_{kj})]^m$.

由于系统中至少有 f_j的一个副本存在,因此数据必定被命中. 这样,命中 f_j只有 3 种可能:在当前结点 n_k或在本地子网格 g_i或在其他子网格中. 因此:

$$P(Eg_{kj})=1-P(El_{kj})-P(EG_{-kj}).$$

针对每一次数据请求,分别统计提供数据服务结点的类型,以数据累积请求为基数,可获得在本地结点、子网格内部、其他子网格的累积平均命中比例,分别记为 $P(local)$,$P(intra)$,$P(inter)$. 若子网格内部用局域网连接,子网格之间用广域网连接,则这些比例说明了文件请求对局域网带宽、广域网带宽的占用比例. 累积平均本地结点命中比例(简称为本地命中比例)为

$$P(local)=\frac{1}{M}\sum_{j=1}^{N}\left[m\frac{\lambda_{hj}}{\lambda}\rho_{hj}+m(G-1)\frac{\lambda_{lj}}{\lambda}\rho_{lj}\right]=\frac{1}{G}\sum_{j=1}^{N}[\lambda_{hj}\rho_{hj}+(G-1)\lambda_{lj}\rho_{lj}]. \tag{1}$$

累积平均其他子网格命中比例(简称为全局命中比例)为

$$P(inter)=\frac{1}{M}\sum_{j=1}^{N}\left[m\frac{\lambda_{hj}}{\lambda}(1-\rho_{hj})^m+(G-1)m\frac{\lambda_{lj}}{\lambda}(1-\rho_{lj})^m\right]=\frac{1}{G}\sum_{j=1}^{N}\left[\lambda_{hj}(1-\rho_{hj})^m+(G-1)\lambda_{lj}(1-\rho_{lj})^m\right]. \tag{2}$$

根据各种命中比例的定义可知,累积平均本地

子网格命中比例(简称为子网格命中比例)为

$$P(intra)=1-P(local)-P(inter). \quad (3)$$

假设访问本地存储空间获取数据的开销是 t_l,访问本地子网格内结点获取数据的开销是 t_g,通过其他子网格的结点获取数据的开销 t_G;其中 $t_l \leqslant t_g \leqslant t_G$.其中,$t_G$指对其他子网格中任意一个结点的访问开销,并假定从其他子网格中任意一个结点访问数据的开销相等.

根据上述累积平均命中比例,可以得到访问任意数据的期望访问延迟:

$$t=P(local)\ t_l+P(intra)\ t_g+P(inter)\ t_G=\frac{1}{G}\sum_{j=1}^{N}\lambda_{hj}[t_g-t_{gl}\rho_{hj}+t_{Gg}\ (1-\rho_{hj})^m]+\frac{1}{G}\sum_{j=1}^{N}(G-1)\lambda_{lj}[t_g-t_{gl}\rho_{lj}+t_{Gg}\ (1-\rho_{lj})^m], \quad (4)$$

其中,$t_{gl}=t_g-t_l$以及 $t_{Gg}=t_G-t_g$.

对所有副本策略需满足以下约束条件,系统中所有数据副本之和不得超过整体存储空间,即 $\sum_{j=1}^{N}[m\rho_{hj}+(G-1)m\rho_{lj}]\leqslant MK$,由于在剩余空间创建副本不会降低系统性能,因此假设所有的存储空间将会占满,化简后可得:

$$\frac{1}{G}\sum_{j=1}^{N}[\rho_{hj}+(G-1)\rho_{lj}]=K. \quad (5)$$

2.2 常用副本策略

根据对相关工作[6-9]的总结,有以下几种比较经典的副本策略:均匀策略、比例策略、平方根策略.

在均匀策略中,不论请求如何,所有数据的副本数目相同,因此副本概率相同,考虑到系统整体空间约束式(5),可得 $\rho_{hj}=\rho_{lj}=K/N$.

在比例策略中,副本概率与请求频率呈正比关系,即 $\rho_{hj}\propto\lambda_{hj}$,$\rho_{lj}\propto\lambda_{lj}$,$j=1,2,\cdots,N$,考虑到空间约束条件,可得:$\rho_{hj}=K\times\lambda_{hj}$,$\rho_{lj}=K\times\lambda_{lj}$.

在平方根策略中,副本概率与请求频率的平方根呈正比关系,即 $\rho_{hj}\propto\sqrt{\lambda_{hj}}$,$\rho_{lj}\propto\sqrt{\lambda_{lj}}$,考虑到空间约束条件,可得:$\rho_{hj}=s\sqrt{\lambda_{hj}}$,$\rho_{lj}=s\sqrt{\lambda_{lj}}$,其中 $s=KG/\sum_{j=1}^{N}[\sqrt{\lambda_{hj}}+(G-1)\sqrt{\lambda_{lj}}]$.

给定一种请求分布,根据每种副本策略中副本概率与请求频率的关系,可以计算出相应的副本分布.

3 副本分布与请求分布的关系

第 1 节假设在 GH_j 的子网格中,数据副本概率要大于 GL_j 的子网格,这仅是一个直观上的假设,本节通过推导证明该假设在平均下载延迟最小的意义上是否成立.另外一个问题是如果系统中请求均匀分布在所有子网格中,即 λ_{hj} 与 λ_{lj} 相同时,数据副本应当如何分布.最后讨论不同文件之间副本分布的关系.

对于第 2.2 节中提到的副本策略而言,GH_j 的子网格中副本概率要大于 GL_j 子网格中副本概率.以比例复制策略为例进行说明.对于文件 f_j,若 $\lambda_{hj}\geqslant\lambda_{lj}$,则 $\rho_{hj}=K\times\lambda_{hj}\geqslant K\times\lambda_{lj}=\rho_{lj}$.相应的,对于文件 f_j 与文件 f_x,若 $\lambda_{lj}\geqslant\lambda_{lx}$,则 $\rho_{lj}\geqslant\rho_{lx}$;若 $\lambda_{hj}\geqslant\lambda_{hx}$,则 $\rho_{hj}\geqslant\rho_{hx}$;可见,若将系统中所有的请求频率进行排序,则对应的副本概率也仍然符合这个次序.

下面讨论最优下载副本策略中是否仍然具有这个特性.根据第 2 节的讨论,如果要使系统中平均下载延迟最小,所对应的优化问题为

$$\min\left\{\frac{1}{G}\sum_{j=1}^{N}\lambda_{hj}[t_g-t_{gl}\rho_{hj}+t_{Gg}\ (1-\rho_{hj})^m]+\frac{1}{G}\sum_{j=1}^{N}(G-1)\lambda_{lj}[t_g-t_{gl}\rho_{lj}+t_{Gg}\ (1-\rho_{lj})^m]\right\}. \quad (6)$$

$$\text{s.t.}\quad \frac{1}{G}\sum_{j=1}^{N}[\rho_{hj}+(G-1)\rho_{lj}]=K,\quad 0\leqslant\rho_{lj}\leqslant\rho_{hj}\leqslant 1.$$

式(6)是一个非线性优化问题,没有显式解.求解上述问题的经典方法是拉格朗日乘子法:

$$H=\frac{1}{G}\sum_{j=1}^{N}\lambda_{hj}[t_g-t_{gl}\rho_{hj}+t_{Gg}\ (1-\rho_{hj})^m]+\frac{1}{G}\sum_{j=1}^{N}(G-1)\lambda_{lj}[t_g-t_{gl}\rho_{lj}+t_{Gg}\ (1-\rho_{lj})^m]+\gamma\left\{\frac{1}{G}\sum_{j=1}^{N}[\rho_{hj}+(G-1)\rho_{lj}]-K\right\}. \quad (7)$$

将式(7)对所有的 ρ_{hj} 求偏导数,可得:

$$\frac{\partial H}{\partial\rho_{hj}}=-\frac{1}{G}\lambda_{hj}[t_{gl}+mt_{Gg}\ (1-\rho_{hj})^{m-1}]+\gamma\frac{1}{G}=0. \quad (8)$$

式(8)化简后可得:

$$\lambda_{hj}[t_{gl}+t_{Gg}m(1-\rho_{hj})^{m-1}]=\gamma. \quad (9)$$

将式(7)对所有的 ρ_{lj} 求偏导数,可得:

$$\frac{\partial H}{\partial\rho_{lj}}=-\frac{G-1}{G}\lambda_{lj}[t_{gl}+mt_{Gg}\ (1-\rho_{lj})^{m-1}]+\gamma\frac{G-1}{G}=0. \quad (10)$$

式(10)化简后可得:

$$\lambda_{lj}[t_{gl}+t_{Gg}m(1-\rho_{lj})^{m-1}]=\gamma. \quad (11)$$

3.1 相同文件

结合式(9)以及式(11)可以得到针对同一个文件 f_j 的高请求频率子网格以及低请求频率子网格间请求频率以及副本概率之间的关系:

$$\lambda_{hj}[t_{gl}+t_{Gg}m(1-\rho_{hj})^{m-1}]=\lambda_{lj}[t_{gl}+t_{Gg}m(1-\rho_{lj})^{m-1}]. \quad (12)$$

由式(12)可知,若 $\lambda_{hj}>\lambda_{lj}$,则

$$t_{gl}+t_{Gg}m(1-\rho_{hj})^{m-1}<t_{gl}+t_{Gg}m(1-\rho_{lj})^{m-1}.$$

化简后可得 $\rho_{hj}>\rho_{lj}$,即对于文件 f_j 请求频率高的子网格内部,副本概率应当高于请求频率低的子网格,这说明在最优复制策略的意义上,本文的假设是成立的.另外,若 $\lambda_{hj}=\lambda_{lj}$,则 $\rho_{hj}=\rho_{lj}$.这意味着若一个文件的请求在系统中均匀分布,则在平均访问延迟最优意义上,其副本应在系统中均匀分布.

3.2 不同文件

采用类似的方法,可以讨论不同文件请求频率以及副本概率之间关系.

结合式(9)以及式(11),可以得到针对文件 f_j 以及文件 f_y 的请求频率高的子网格中请求频率与副本概率的关系:

$$\lambda_{hj}[t_{gl}+t_{Gg}m(1-\rho_{hj})^{m-1}]=\lambda_{hy}[t_{gl}+t_{Gg}m(1-\rho_{hy})^{m-1}]. \quad (13)$$

若 $\lambda_{hj}\leqslant\lambda_{hy}$,可以得出 $\rho_{hj}\leqslant\rho_{hy}$;若 $\max\lambda_{hj}=\lambda_{hy}$,则有 $\max\rho_{hj}=\rho_{hy}$,即文件 f_y 的请求频率高的子网格中副本概率为系统中最高.

相应地,可以得出文件以及文件的请求频率低的子网格中请求频率与副本概率之间的关系:

$$\lambda_{lj}[t_{gl}+t_{Gg}m(1-\rho_{lj})^{m-1}]=\lambda_{lx}[t_{gl}+t_{Gg}m(1-\rho_{lx})^{m-1}]. \quad (14)$$

若 $\lambda_{lj}\geqslant\lambda_{lx}$,则 $\rho_{lj}\geqslant\rho_{lx}$,假设对所有文件的请求频率中有 $\min\lambda_{lj}=\lambda_{lx}$,则 $\min\rho_{lj}=\rho_{lx}$,即文件 f_x 的请求频率低的子网格中副本概率为系统中最低.

下面讨论文件的请求频率高的子网格以及文件请求频率低的子网格请求频率以及副本概率的关系,对于两个文件 f_j 与 f_x,根据式(9)以及式(11)有:

$$\lambda_{hj}[t_{gl}+t_{Gg}m(1-\rho_{hj})^{m-1}]=\lambda_{lx}[t_{gl}+t_{Gg}m(1-\rho_{lx})^{m-1}]. \quad (15)$$

若 $\lambda_{hj}\geqslant\lambda_{lx}$,则 $\rho_{hj}\geqslant\rho_{lx}$;否则,若 $\lambda_{hj}\leqslant\lambda_{lx}$,则 $\rho_{hj}\leqslant\rho_{lx}$.

综合上面的讨论可以得出,将系统中所有文件的子网格请求频率按照降序进行排列,$\lambda_{hy}\geqslant\cdots\geqslant\lambda_{hj}\geqslant\cdots\geqslant\lambda_{lj}\geqslant\cdots\geqslant\lambda_{ly}\geqslant\cdots\geqslant\lambda_{hx}\geqslant\cdots\geqslant\lambda_{lx}$,则相应的副本概率也会按照相同的次序进行排列,即 $\rho_{hy}\geqslant\cdots\geqslant\rho_{hj}\geqslant\cdots\geqslant\rho_{lj}\geqslant\cdots\geqslant\rho_{ly}\geqslant\cdots\geqslant\rho_{hx}\geqslant\cdots\geqslant\rho_{lx}$.可见,在优化策略中,若某个子网格对文件的请求频率高,则该子网格对该文件拥有较高的副本概率.

4 求解最优下载副本策略

假设对所有文件的请求频率中有 $\min\lambda_{lj}=\lambda_{lx}$,则 $\min\rho_{lj}=\rho_{lx}$,考虑到约束条件式(5)可知,$\rho_{lx}\leqslant K/N$;否则,如果 $\rho_{lx}>K/N$,则:

$$\frac{1}{G}\sum_{j=1}^{N}[\rho_{hj}+(G-1)\rho_{lj}]>\frac{1}{G}\sum_{j=1}^{N}\left[\frac{K}{N}+(G-1)\frac{K}{N}\right]=K.$$

可见,这样所有的副本将会超出系统整体存储空间的上限,与式(5)矛盾,因此系统中请求频率最低的文件的副本概率上限为 K/N.

根据式(14)、式(15)可以知道所有文件的副本概率 ρ_{lj} 以及 ρ_{hj} 可以表示成 ρ_{lx} 的函数.

令:

$$h(j,\rho_{lx})=\lambda_{lx}[t_{gl}+t_{Gg}m(1-\rho_{lx})^{m-1}]/(\lambda_{hj}t_{Gg}m)-t_{gl}/(t_{Gg}m),$$

则 $\rho_{hj}=1-[h(j,\rho_{lx})]^{1/(m-1)}$.

令:

$$l(j,\rho_{lx})=\lambda_{lx}[t_{gl}+t_{Gg}m(1-\rho_{lx})^{m-1}]/(\lambda_{lj}t_{Gg}m)-t_{gl}/(t_{Gg}m),$$

则 $\rho_{lj}=1-[l(j,\rho_{lx})]^{1/(m-1)}$.

下面分析 ρ_{hj},ρ_{lj} 的特征,将 ρ_{hj},ρ_{lj} 分别对 ρ_{1x} 求导可得:

$$\frac{d\rho_{hj}}{d\rho_{lx}}=\frac{\lambda_{lx}}{\lambda_{hj}}(1-\rho_{lx})^{m-2}[h(j,\rho_{lx})]^{\frac{1}{m-1}-1}>0,$$

$$\frac{d\rho_{lj}}{d\rho_{lx}}=\frac{\lambda_{lx}}{\lambda_{lj}}(1-\rho_{lx})^{m-2}[l(j,\rho_{lx})]^{\frac{1}{m-1}-1}>0.$$

令 $f(\rho_{1x})=\frac{1}{G}\sum_{j=1}^{N}[\rho_{hj}+(G-1)\rho_{lj}]-K$,因此有:

$$\frac{df(\rho_{lx})}{d\rho_{lx}}=\frac{1}{G}\sum_{j=1}^{N}\left[\frac{d\rho_{hj}}{d\rho_{lx}}+(G-1)\frac{d\rho_{lj}}{d\rho_{lx}}\right]>0.$$

可以看出 $f(\rho_{lx})$ 是一个严格单调递增函数,于是,若 $f(0)\leqslant 0$ 且 $f(K/N)\geqslant 0$,则 $f(\rho_{lx})$ 在其定义域内有解.本文讨论方程有解的情况,若无解,则本文方法失效,说明本文的优化问题在可行空间内无解,该问题留待后续工作进行研究.下面的讨论中假设方程有解,从而优化问题可以求得最优解.对方程 $f(\rho_{lx})=0$ 采用一维搜索优化算法可以进行求解.本文采用二分法(bisection method)确定 ρ_{lx} 的数

值.下面给出本文中用到的求解 $f(\rho_{lx})=0$ 的算法,假设 $f(\rho_{lx})=0$ 的解为 ρ_{lx}^*:

1) 若 $f(0)>0$ 或 $f(K/N)<0$,则方程无解,返回;

2) 设定精度 δ,令 $x_l=0, x_r=K/N$;

3) 若 $|x_r-x_l|<\delta$,则 $\rho_{lx}^*=(x_r+x_l)/2$,返回;

4) 令 $x_m=(x_l+x_r)/2$;

5) 计算 $f(x_m)$,若 $f(x_m)=0$,则 $\rho_{lx}^*=x_m$,返回;若 $f(x_m)<0$,则 $x_l=x_m$,否则 $x_r=x_m$,转第 3 步.

平均来讲,二分法每次可以将搜索区间缩小一半,收敛速度为 1/2,并非是最有效的方法,在后续工作中将尝试其他的搜索优化方法进行求解.

求解 ρ_{lx} 后就可确定其他 ρ_{hj} 以及 ρ_{lj} 的数值,从而计算出最优的副本分布,所求最优分布为

$$\begin{cases}\rho_{hj}=1-\sqrt[m-1]{h(j,\rho_{lx}^*)},\\ \rho_{lj}=1-\sqrt[m-1]{l(j,\rho_{lx}^*)}, j\neq x,\\ \rho_{lx}=\rho_{lx}^*.\end{cases}\tag{16}$$

本文将其称为数据请求具有分组特性的最优下载副本策略(optimized downloading replication strategy for clustered demands,ODRS-C).与其他几种策略相比较,ODRS-C 需要求解非线性方程,计算复杂度最高.

5 测试结果及性能分析

5.1 实验设置

在后续的模拟实验中,采用以下参数配置.结点总数 $M=1000$,子网格规模 $m=10$,结点容量 $K=10$,文件数目 $N=100$.每个文件的大小为 1 GB,子网格内部用于传输数据的带宽为 100 Mbps,因此 $t_g=80$ s.子网格之间用于传输数据的带宽为 10 Mbps,因此 $t_G=800$ s.由于本地结点命中时不引起网络开销,因此 $t_l=0$ s.许多针对网络应用的测试工作表明,文件请求分布可用 zipf 分布[14]刻画.因此,本文关于文件请求的分布以 zipf 分布为例,对系统进行分析与模拟,选用 zipf 分布参数 $\alpha=0.6$,关于其他参数的测试结果略去.根据文献[11]的测试结果,请求分组主要发生在请求频度比较低的一些文件上,因此本文的测试中,最热门的 20 个文件 f_j 的请求在系统中均匀分布,即 $c_j=1\%$, $j=1, 2, \cdots, 20$.对于其他文件,有 10 种分组模式,聚集比例从 1%到 10%.对于文件 f_j,任意一个子网格属于 GH_j 的概率相同.从系统全局来看,每个文件的整体请求频度仍符合 zipf 分布.

针对第 2.2 节以及第 4 节的复制策略,首先根据相应的副本分布计算出每个文件的副本数目,然后按照每个结点拥有某个文件的概率随机分配到系统的各个结点.一个文件的副本最多在一个结点上出现一次,若结点空间已满,则寻找其他空闲结点分配副本.分配数据副本之后,每个结点按照设定的请求分布发出对文件的请求,并记录每次请求的命中类型:本地命中、本地子网格命中、其他子网格命中,分别对这些命中次数进行统计.由于系统采用随机分配的方法,因此每次运行分配算法时,结点中数据副本的分配结果不同.针对一种系统配置,运行 100 次分配算法,共产生 100 种分配结果.针对每种分配结果,运行 1000 次迭代,其中一次迭代中,每个结点平均发出 1 次文件请求.

在数据网格中广泛采用缓存替换策略来管理存储空间,本文针对一种经典的缓存替换策略 LRU 进行实验,测试其在请求分组特性下的性能.每个结点采用 LRU 算法管理自己的空间,若存储空间满,则采用 LRU 算法替换数据.在该实验中,所产生的请求次数与上面针对副本策略相同.

将所有文件请求的命中类型进行统计,除以所有的请求次数,就可获得第 2.1 节定义的累积平均命中比例,从而可计算出一次文件请求所需的平均延迟.

5.2 命中比例

在后续图 1~4 中,P 代表比例复制策略,S 代表平方根复制策略,O 代表最优下载复制策略,U 代表均匀复制策略,LRU 代表 LRU 缓存替换策略.

图 1 给出了在第 5.1 节参数设置下,实验测出的各种策略的本地命中比例.可以看出,请求分组对于均匀策略的性能没有影响,在各种分组模式下,均匀策略的性能几乎维持不变.对于其他 4 种策略,与

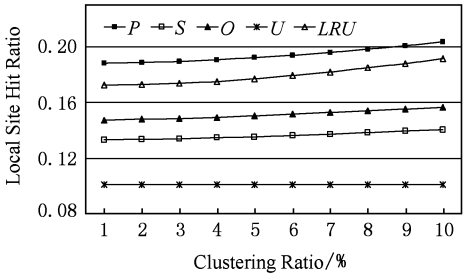

Fig. 1 Local site hit ratio trend.

图 1 本地命中比例

请求均匀分布的模式相比,请求分组造成了本地命中比例的增加;此外,随着请求聚集比例的升高,本地命中比例在增加.其中,比例复制策略拥有最高的本地命中比例,*LRU* 策略次之.

图 2 给出的是子网格命中比例,即局域网带宽开销,可以看出,均匀策略仍然与请求分组无关.平方根策略所需要的局域网带宽最多,优化策略次之;比例策略与 *LRU* 策略的子网格命中比例比较接近,所需要的局域网带宽最少.

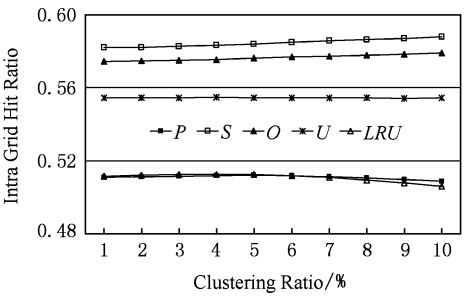

Fig. 2 Intra grid hit ratio trend.

图 2 子网格命中比例

图 3 给出的是全局命中比例,也就是广域网带宽占用情况.与前面类似,均匀策略几乎与请求分组无关.在所有策略中,均匀策略所需广域网带宽最多,优化策略所占用的广域网带宽最少;比例复制策略尽管拥有最高的本地命中比例,但是所需消耗的广域网带宽要高于平方根策略以及优化策略.此外,LRU 所需广域网带宽也要高于平方根策略以及优化策略.

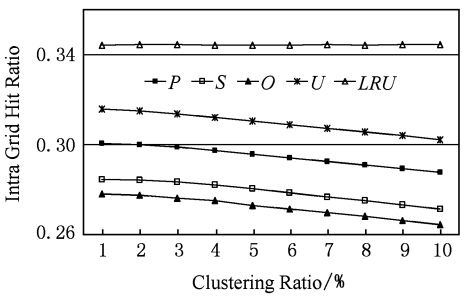

Fig. 3 Inter grid hit ratio trend.

图 3 全局命中比例

可见,从系统整体角度讲,单纯提高本地命中比例并不一定导致性能最优,而是需要综合考虑各种命中的比例关系.除均匀策略外,所有分组模式所导致的广域网带宽占用要少于请求均匀分布的情况;随着聚集比例的增加,广域网带宽开销在逐渐下降.可见,在本文的测试中,请求分组导致系统性能改善.

5.3 平均下载延迟

图 4 给出了平均下载延迟的趋势.与命中比例趋势类似,均匀策略的平均延迟与请求分组几乎无关.对于其他几种策略而言,变化趋势与图 3 相同,这是由于在决定平均访问延迟的因素中,广域网带宽起着主导作用.可以看出,优化策略所对应的平均下载延迟最小,这说明优化模型所给出的分布达到了预期目标.

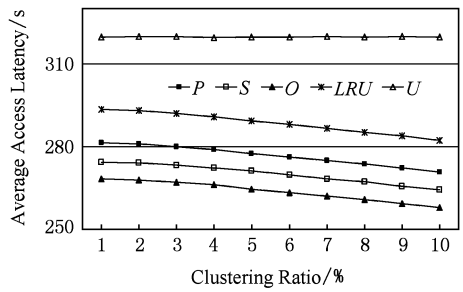

Fig. 4 Average access latency trend.

图 4 平均访问延迟

6 结束语

本文以文件下载的期望延迟作为优化目标,通过对请求具有分组模式的联合型数据网格中用户请求分布与副本分布之间关系的研究,建立了数据访问期望延迟的优化模型,进行求解后,得出一个最优下载复制策略(ODRS-C).与常用的均匀复制策略、比例复制策略、平方根复制策略以及 LRU 缓存策略相比较,根据 ODRS-C 所对应的数据副本分布在系统中创建副本,可以使得文件请求对广域网的带宽占用最小,从而使得一个文件的期望下载时间最小化.本文工作建立在理论分析与模拟测试基础上,因此,工作需要根据该分布在实际系统中测试,并与理论分析结果以及模拟结果进行比较.另外,本文所基于的联合型数据网格是完全对称的结构,结点的存储空间、文件的大小、子网格规模大小相同,而实际系统会呈现出一定的异构性.因此,需要对具有请求分组特性的异构系统的副本优化问题进行研究,从而获得更加一般的结果.

参 考 文 献

[1] Wang Yijie, Xiao Nong, Ren Hao, *et al*. Research on key technology in data grid [J]. Journal of Computer Research and Development, 2002, 39(8): 943-947 (in Chinese)
(王意洁, 肖侬, 任浩, 等. 数据网格及其关键技术研究[J]. 计算机研究与发展, 2002, 39(8): 943-947)

[2] Foster I, Kesselman C. The Grid: Blueprint for a New Computing Infrastructure [M]. Beijing: China Machine Press, 2005: 391-429

[3] Bell W H, Cameron D G, Carvajal-Schiaffino R, *et al*. Evaluation of an economy-based file replication strategy for a data grid [C] //Proc of the 3rd IEEE/ACM Int Symp on Cluster Computing and the Grid. Los Alamitos, CA: IEEE Computer Society, 2003

[4] Ekow O, Doron R, Alexandru R. Optimal file-bundle caching algorithms for data-grids [C] //Proc of the 2004 ACM/IEEE Conf on Supercomputing. Los Alamitos, CA: IEEE Computer Society, 2004

[5] Iamnitchi A, Doraimani S, Garzoglio G. Filecules in high-energy physics: Characteristics and impact on resource management [C] //Proc of the 15th IEEE Int Symp on High Performance Distributed Computing. Los Alamitos, CA: IEEE Computer Society, 2006: 69-80

[6] Cohen E, Shenker S. Replication strategies in unstructured peer-to-peer networks [C] //Proc of the Conf on Applications, Technologies, Architectures, and Protocols for Computer Communications. New York: ACM, 2002: 177-190

[7] Lv Q, Cao P, Cohen E, *et al*. Search and replication in unstructured peer-to-peer networks [C] // Proc of the 16th Int Conf on Supercomputing. New York: ACM, 2002: 84-95

[8] Tewari S, Kleinrock L. Proportional replication in peer-to-peer network [C] //Proc of the 25th IEEE Int Conf on Computer Communications. Los Alamitos, CA: IEEE Computer Society, 2006

[9] Tewari S, Kleinrock L. Optimal search performance in unstructured peer-to-peer networks with clustered demands [J]. IEEE Journal on Selected Areas in Communications, 2007, 25(1): 84-95

[10] Iamnitchi A, Ripeanu M, Foster I. Small-world file-sharing communities [C] //Proc of the 23rd IEEE Int Conf on Computer Communications. Los Alamitos, CA: IEEE Computer Society, 2004

[11] Handurukande S B, Kermarrec A -M, Fessant F Le, *et al*. Peer sharing behaviour in the eDonkey network, and implications for the design of server-less file sharing systems [C] //Proc of the ACM SIGOPS EuroSys Conference. New York: ACM, 2006

[12] Klemm A, Lindemann C, Vernon M K, *et al*. Characterizing the query behavior in peer-to-peer file sharing systems [C] // Proc of ACM Internet Measurement Conference (IMC). New York: ACM, 2004

[13] Jiang J, Yang G. An optimal replication strategy for data grid systems [J]. Frontiers of Computer Science in China, 2007, 1(3): 338-348

[14] Breslau L, Cao P, Fan L, *et al*. Web caching and Zipf-like distributions: Evidence and implications [C] //Proc of the 18th IEEE Int Conf on Computer Communications. Los Alamitos, CA: IEEE Computer Society, 1999: 126-134

Jiang Jianjin, born in 1972. Ph. D. His main research interests are grid computing and peer to peer system.

姜建锦,1972 年生,博士,主要研究方向为网格计算、对等网络.

Yang Guangwen, born in 1963, Ph. D., professor and Ph. D. supervisor. His main research interests include grid computing, parallel and distributed processing, and algorithm design and analysis.

杨广文,1963 年生,博士,教授,博士生导师,主要研究方向为网格计算、并行与分布处理、算法设计与分析.

Research Background

This Work is supported by the National Natural Science Foundation of China (90412006, 90412011, 60573110, 90612016, and 60673152), the National Key Basic Research Project of China (2004CB318000 and 2003CB317007), and the National High Technology Development Program of China (2006AA01A101, 2006AA01A108, and 2006AA01A111). In distributed systems including data grid, wide area network bandwidth is still a bottleneck for system performance. By means of efficient replication strategy, wide area network bandwidth requirement can be decreased. Ultimately, average access latency can be decreased too. Most of previous works about replication strategy assume that data request is uniform in system. However, according to some recent traces about data grid and peer to peer systems, data request and replica distribution exhibit clustering properties. Some related work investigates what is the optimal way to replicate data in unstructured peer to peer system when request exhibits clustering properties. The metrics they consider are search performance and query load. This paper investigates download performance of data grid when request exhibits clustering properties. We first give the formal model of replication strategies in federated data grid system. The performance metrics include cumulative hit ratios and average access latency. Then we investigate what is the optimal way to replicate data with the objective of minimizing average access latency when request exhibits clustering properties. We solve the optimization model and get an optimization downloading replication strategy for clustering demands. Simulation results validate the effectiveness of optimal strategy. Compared with some popular strategies, the optimal strategy has some advantages of lower wide area network bandwidth requirement and lower average access latency.