WWW 集群服务器的数据副本分布方式研究*

沈海华，陈世敏，沈美明，郑纬民

(清华大学 计算机科学与技术系，北京 100084)

E-mail: panda@est4.cs.tsinghua.edu.cn

http://www.tsinghua.edu.cn

摘要：为了有效地提高 WWW 服务器的吞吐能力、反应速度和可扩展性，国际上许多著名站点纷纷转向采用 WWW 集群服务器来替代原有的单一主机服务器.采用不同副本分布方式的 WWW 集群服务器，其数据可靠性也有所不同.对不同数据副本分布方式进行探讨，同时，论证了最优副本分布方案.

关键词：WWW 集群服务器；数据副本；副本分布方式；可靠性

中图法分类号：TP393 **文献标识码**：A

随着计算机网络的发展和普及，网上访问的人数和访问频度不断增加，迅速增长的服务流量给 WWW 服务器的处理能力提出了越来越高的要求，尤其在一些热门站点更是如此.为了提高服务器的吞吐能力和反应速度，人们不得不升级或更换服务器.与采用价格昂贵的高性能主机或 SMP 计算机相比，集群服务器以其高性能、可扩展性和低成本脱颖而出.在一些关键性的领域，如金融、电子商务等，都要求 24 小时基本不间断的高可靠性服务，能否解决好高可靠性问题，成为 WWW 集群服务器能否适用于这些服务的关键.

在 WWW 集群中，为了提供结点数据的可靠性，需要将数据保持多个副本.对于双机系统，副本的分布是直截了当的.但是，WWW 集群有多个结点，副本分布可以存在多种方式，例如，Novell 和 Vinca 公司开发的 StandbyServer 中就有“一到一”、“一到多”、“多到一”等多种副本分布方式[1~3].下面，我们将对副本分布方案进行分析和研究，找出最优副本分布方案，并在理论加以证明.

1 保持两个副本时副本分布的可靠性分析

问题 集群系统有 N 个结点($N\geqslant 2$)，每个结点有相同的存储容量.所有数据都在不同结点上有两个副本，同时，所有结点都只保持两个数据副本.求：使整个系统正确提供全部数据可靠性最高的副本分布方式.

本节以下讨论都是针对上述问题进行的，“副本分布方式”若不作特殊声明，都是指满足上述问题前提要求的副本分布方式.使用图来表示副本的分布：

图 $G=(V,E)$，$V=\{$集群的结点$\}$，$E=\{(i,j)\mid i$ 与 j 有同一数据的副本，$i,j\in V\}$.例如，$N=4$ 时可以有两种副本分布方式，如图 1 所示.图 2 将这两种分布表示成图的形式.

* **收稿日期**：1999-09-20；**修改日期**：2000-01-03

基金项目：国家 863 高科技发展计划资助项目(863-306-ZT01-03-1)

作者简介：**沈海华**(1971－)，女，浙江杭州人，博士生，主要研究领域为并行/分布计算机系统，高可用性；**陈世敏**(1973－)，男，北京人，硕士，主要研究领域为并行/分布计算机系统；**沈美明**(1938－)，女，江苏吴县人，教授，博士生导师，主要研究领域为并行/分布计算机系统；**郑纬民**(1946－)，男，浙江宁波人，教授，博士生导师，主要研究领域为并行/分布计算机系统.

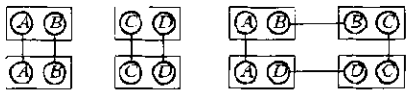

Fig. 1 Data copy distribution patterns when N equals to 4

图1 4个结点的集群副本分布的可能

Fig. 2 Describe data copy distribution patterns when N equals to 4 using graphs

图2 以图表示的4个结点的副本分布可能方式

引理 1. 图 G 是由互不相交的环构成的. 即 $V=V_1\cup\ldots\cup V_k, E=E_1\cup\ldots\cup E_k, k\geqslant 1$. 任取 $i,j=1,\ldots,k$ 且 $i\neq j$, 有 $V_i\cap V_j=\varphi, E_i\cap E_j=\varphi$, 并且图 G 的每个子图 $G_i=(V_i,E_i)$ 都是一个环.

证明: 由于每个结点保存且仅保存两个副本, 而且这两个副本是不同的. 所以, 图 G 中每个顶点的度为 2. 那么可以使用下述算法进行构造性证明:

i) $A\leftarrow V, k\leftarrow 1$

ii) 当 $A\neq\varphi$ 时, 做

iii) $\forall u\subset A, V_k=\{u\}$

iv) 当 u 存在相邻的顶点 $w\in A-V_k$ 时, 做

v) $V_k\leftarrow V_k\cup\{w\}, u\leftarrow w$

vi) $A\leftarrow A-V_k, k\leftarrow k+1$

vii) V_i 构成 $G_i, i=1,\ldots,k$

算法由双层循环构成. 内层循环找到一个环的所有顶点, 外层循环找到每个环. 开始时, 算法步骤 iii)从图 G 的顶点中任意取出一个 u, 由于 u 的度为 2, 所以 u 必有邻点存在. 将 u 的一个邻点 w 放入 V_1 中. 因为 w 的度为 2, 所以除了 (u,w) 边之外, w 必有其他邻边. 可以继续步骤 iv)和 v)的循环. 设这个循环依次找到并加入 V_1 的点为 $u_1,u_2,\ldots,u_m$. 考虑 $u=u_m$ 的情形, 由于 u_m 的度为 2, 所以除了 (u_{m-1},u_m) 之外, u_m 还有邻边 (u_m,w). 由于 $u_2,\ldots,u_m$ 均已经使用了两条边, 所以 $w=u_1$ 或者 $w\notin\{u_1,\ldots,u_m\}$. 如果前者成立, 结束 V_1 循环, $V_1=\{u_1,\ldots,u_m\}$ 构成一个环. 如果后者成立, 则将 w 加入 V_1 继续循环. 但是图 G 的顶点数目是有限的, 所以在某次循环时会有 $w=u_1$ 成立. 这样就找到一个环. 如果图 G 还有其他顶点, 那么继续外部循环, 同理, 找到每一个环. □

于是, 我们先集中考虑单环的情形.

引理 2. 设单个结点正常工作的概率为 p, n 个结点的单环正常提供全部数据的概率为 R_n, 则 $R_n-pR_{n-1}+(1-p)pR_{n-2}$ 对 $n\geqslant 5$ 成立.

证明: 设在单链 $-(0)-(1)-\ldots-(n)-(n+1)-$ 中, 在结点 0 和结点 $n+1$ 正常工作的情况下, 正常提供全部数据的概率为 C_n. 我们先求 C_n 的递推式.

分情况: 当结点 1 正常时, 结点 2 到 n 正常提供全部数据的概率为 C_{n-1}; 当结点 1 失效时, 结点 2 必须正常(否则, 结点 1 与结点 2 之间的公共副本失效), 结点 3 到 n 正常提供全部数据的概率为 C_{n-2}. 有 $C_n=pC_{n-1}+(1-p)pC_{n-2}, n\geqslant 3$; 因为有 $C_0=1, C_1=1, C_2=1-(1-p)^2$, 所以

$$C_n=pC_{n-1}+(1-p)pC_{n-2},\quad n\geqslant 2, \tag{1}$$

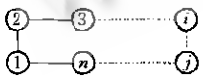

Fig. 3 Circle with n node

图3 n 个结点的环

对于 n 个结点的环, 如图 3 所示, 当结点 1 正常工作时, 结点 2 到 n 正常提供全部数据的概率为 C_{n-1}; 当结点 1 失效时, 结点 2 和结点 n 必须正常工作, 结点 3 到 $n-1$ 正常提供全部数据的概率为 C_{n-3}. 所以有

$$R_n=pC_{n-1}+(1-p)p^2C_{n-3},\quad n\geqslant 3, \tag{2}$$

则组合式(1)和式(2)有

$$R_n-pR_{n-1}-(1-p)pR_{n-2}=0,\quad n\geqslant 5,$$

即

$$R_n-pR_{n-1}+(1-p)pR_{n-2},\ n\geqslant 5. \tag{3}$$

□

引理 3. $R_n\leqslant R_{n-1}, n\geqslant 3$.

证明:由式(1)有 $C_n\geqslant pC_{n-1}$,对于 $n\geqslant 2$ 成立,易验算 $C_1\geqslant pC_0$,所以

$$C_n\geqslant pC_{n-1},\ n\geqslant 1.$$

又 $C_n-C_{n-1}=pC_{n-1}+(1-p)pC_{n-2}-C_{n-1}-(1-p)(pC_{n-2}-C_{n-1})\leqslant 0, n\geqslant 2$. 因为 $C_1=C_0$,所以

$$C_{n-1}\geqslant C_n\geqslant pC_{n-1},\quad n\geqslant 1. \tag{4}$$

由式(2)有 $R_n-R_{n-1}=p(C_{n-1}-C_{n-2})+(1-p)p^2(C_{n-3}-C_{n-4})\leqslant 0, n\geqslant 4$.

$$R_2=1-(1-p)^2=2p-p^2,$$

$$R_3=pC_2+(1-p)p^2C_0=p(2p-p^2)+(1-p)p^2=p2(3-2p),$$

$$R_3-R_2=p^2(3-2p)-(2p-p^2)=p(-2+4p-2p^2)=-2p(1-p)^2\leqslant 0.$$

所以

$$R_n\leqslant R_{n-1}, n\geqslant 3. \tag{5}$$

□

定理. 对于 $N(N\geqslant 2)$ 个结点的集群,所有数据都在不同结点上有两个副本,同时,所有结点都只保持两个副本. 当 $N=2n$ 时,组成 n 个双结点环具有最高的数据可靠性,$\max R_N=R_2^n=[1-(1-p)^2]^n$. 当 $N=2n+1$ 时,组成 n 个双结点环和 1 个三结点环的可靠性最高,$\max R_N=R_3R_2^n=p^2(3-2p)\cdot[1-(1-p)^2]^n$.

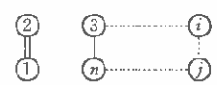

Fig. 4 A circle with n nodes is divided into a circle with two nodes and a circle with $n-2$ nodes

图4 n 顶点单环分解为一个双顶点单环和一个 $n-2$ 个顶点单环

证明:①将图 3 中的单环分解为一个双顶点单环和一个 $n-2$ 个顶点单环的形式,如图 4 所示.

新结构的系统提供全部数据的可靠性为 $R'_n=R_2\times R_{n-2}$,则由引理 2 和引理 3 有,

$$R'_n-R_n=pR_{n-1}+(1-p)pR_{n-2}-(2p-p^2)R_{n-2}=pR_{n-1}+(-p)R_{n-2}$$

$$=p(R_{n-1}-R_{n-2})\geqslant 0,\ n\geqslant 5.$$

由式(1)和式(3)可求得,$R_4=p^2(2-p^2)$,有 $R_2^2-R_4=2p^2(1-p)^2\geqslant 0$.

所以,结点数大于等于 4 的单环分解后可靠性非减.

②由引理 1,图 C 是由互不相交的环构成的. 对于任意一个环,都可以按①分解,而保持系统可靠性非降. 最后得到的图由双结点和三结点的环组成.

③对于两个三结点环,将它们组成 3 个双结点环:

$$R_3^3-R_3^2=p^3(2-p)^3-p^4(3-2p)^2=p^5(1-p)^2(8-5p)\geqslant 0,$$

保持系统可靠性非降(易验证:$R_2-R_1^2\geqslant 0, R_3-R_1R_2\geqslant 0$. 不能继续分解为单个结点的形式). 所以,通过②分解和③重组,最终将得到至多一个三结点环(当 N 为奇数时)和 $N \bmod 2$ 个双结点环,使正常提供全部数据的可靠性达到最大值. 而且有

当 $N=2n$ 时,$\max R_N=R_2^n=[1-(1-p)^2]^n$;

当 $N=2n+1$ 时,$\max R_N=R_3R_2^n=p^2(3-2p)[1-(1-p)^2]^n$. □

2 副本分布方式综合分析

上一节我们讨论了保持两个副本时保证系统正确提供数据的可靠性最高的方式——双环方

式.表1给出了在结点正常工作的概率为0.99或0.995的情况下,分别取双结点组数为4、8、16、32、64时计算的结果.为了使R_N不低于0.999,在$p=0.995$的情况下,可以使用64个结点.可见,这种分布方式在成本、可扩展性、可管理性和实现的复杂性这些因素的衡量下是较好的,是适于集群系统的.

Table 1 Reliability that N nodes can provide data correctly when 2 data replicas are remained

表1 保持两个副本时N个结点正常提供全部数据的可靠性

	$g=4,N=8$	$g=8,N=16$	$g=16,N=32$	$g=32,N=64$	$g=64,N=128$
$p=0.99$	0.999 6	0.999 2	0.998 4	0.996 8	0.993 6
$p=0.995$	0.999 9	0.999 8	0.999 6	0.999 2	0.998 4

进一步,如果需要在一定的可靠性下采用更多的结点,那么可以推广双环方式的结论,考虑保持m个副本的情形.可以提出如下的分布方式:

对于$N=g\times m$个结点($g\geq 1, m\geq 2$),要求每个数据有m个副本分布在m个不同的结点上,每个结点保持m个不同的副本.可以将集群分成每组m个结点的多个复制组,同组结点互相保持副本.这时,$R_N=[1-(1-p)^m]^g$.

与双环方式相似,在这种方式中结点是同等的,分布规则符合可管理性、可扩展性的要求,结点组之间没有联系,相对于保持m个副本的其他分布来说,实现的复杂性较小.而且它的可靠性也得到进一步提高.表2给出了$m=3$时的计算结果.在$p=0.99$的情况下,192个结点的大集群也可以有0.999 936的可靠性.

Table 2 Reliability that N nodes can provide data correctly when 3 data replicas are remained

表2 保持3个副本时N个结点正常提供全部数据的可靠性

	$g=4,N=12$	$g=8,N=24$	$g=16,N=48$	$g=32,N=96$	$g=64,N=192$
$p=0.99$	0.999 996	0.999 992	0.999 984	0.999 968	0.999 936
$p=0.995$	0.999 999 5	0.999 999	0.999 998	0.999 996	0.999 992

除了可靠性因素之外,副本分布方式的选择还受到集群的成本、可扩展性、可管理性和实现的复杂性等因素[4]的影响.

以"多到一"副本分布方式为例.如果将集群分组,每组采用"多到一"方式,由于保持副本的结点需要较大的存储容量,为了达到同样的单机可靠性,机器价格必定更加昂贵,这必将增加集群的成本.如果采用价格与普通结点相当的机器,那么保持副本的结点的可靠性就低于普通结点,成为系统的薄弱环节[6].

可见,集群中所有的结点都应同等对待.特殊的结点可能会引起成本上升,从而限制了系统的可扩展性.

在保持一定可靠性的前提下,系统可以采用足够多的结点,保持可扩展性.

副本分布应可以按照一定的规则进行.在给定结点数目N后,就可以按照这种规则容易地给出副本的分布.而且当结点数目增加时,新的分布基本不需要改变已有的副本分布.这样才符合可管理性和可扩展性的要求[6].

此外,副本的分布规则越简单,结点之间的耦合越少,实现也就越容易.在上节得到的最佳副本分布中,所有的结点都被同等对待,并保持两个副本,按照双结点环的方式进行副本的分布.在新增加结点时,只需两两配对即可,原来的副本分布可以不做任何改变.不同的双环间不需要联系,从而简化了实现.

3 结束语

提供数据的高可靠性作为衡量WWW集群服务器服务质量的指标正变得越来越重要,当前国

际上通行的提高数据可靠性的方式是提供数据冗余(即为一个数据保存多个副本)[7]. 限于集群服务器的成本,副本分布不可能过多. 本文对 WWW 集群服务器的数据副本分布方式进行了分析,并在给定的前提下提出并论证了最优副本分布方式. 目前,该理论已在我们开发的集群服务器 TH-PARAWEB 上得到实践.

References:

[1] Novell Inc., StandbyServerTM for NetWare/intraNetWare, White Paper, 1998. http://www.nevell.com/products/clusters/sbs.

[2] Novell Inc., StandbyServerTM Many-to-One for NetWare/intraNetWare, White Paper, 1998. http://www.nevell.com/products/clusters/sbsmto/sbsmto-up.html.

[3] Vinca, Corporation. StandbyServerTM Many-to-One for NetWare User's Guide, 1998. http://www.vinea.com/products/sbsht.

[4] Liu, Pin. The Foundation of Reliability Engineering. Beijing: China Computation Press, 1995. 1~7.

[5] Guerraoui, R., Schiper, A. Software-Based replication for fault tolerance. Computer, 1997,30(4):68~74.

[6] Short, R., Gamache, R., Vert, J., *et al*. Windows NT clusters for availability and scalability. In: Werner, R., ed. Proceedings of the 42nd Annual IEEE International Computer Conference, San Jose, CA: IEEE Piscataway, 1997. 8~13.

[7] Watts, D., Credle Rufusm, Jr. Pelles Joao, *et al*. Clustering and high availability guide for IBM netfinity and IBM PC servers. SG24 4858 00, 1997. http://www.redbooks.ibm.com/SG244858/wolfpack.htm.

Research on Distribution of Data Copy for WWW Server Clusters*

SHEN Hai-hua, CHEN Shi-min, SHEN Mei-ming, ZHENG Wei-min

(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*)

E-mail: panda@est4.cs.tsinghua.edu.cn

http://www.tsinghua.edu.cn

Abstract: In order to improve the throughput, response speed and scalability of Web servers efficiently, many famous Web sites have thrown away single servers and turned to Web server clusters. Web server clusters with different distribution modes of data copy have different data availability levels. After discussing many data copy distribution modes, an optimized data copy distribution is proved in this paper.

Key words: WWW server cluster; data copy; data copy distribution; reliability

* Received September 20, 1999; accepted January 3, 2000

Supported by the National High Technology Development Program of China under Grant No. 863-306-ZT01-03-01