分布式存储系统的可靠性研究

张 薇[1,2]，马建峰[2]，杨晓元[1]

(1. 武警工程学院 电子技术系，陕西 西安 710086；
2. 西安电子科技大学 计算机网络与信息安全教育部重点实验室，陕西 西安 710071)

摘要：通过对分布式存储系统体系结构的研究，认为数据服务的可靠性由时间、节点失效概率密度函数、数据分离算法及存储策略这 4 个因素影响. 在此基础上，结合可靠性理论，用概率方法构造了存储系统的可靠性模型，根据模型可以对给定系统的可靠性进行预测，并据此制定存储策略，从而将可靠性问题在系统设计阶段解决，并使数据服务的可靠性保持在较高水平.

关键词：存取结构；失效概率；存储策略

中图分类号：TP309 **文献标识码**：A **文章编号**：1001-2400(2009)03-0480-06

Reliability of distributed storage systems

ZHANG Wei[1,2], *MA Jian-feng*[2], *YANG Xiao-yuan*[1]

(1. Engineering Institute of the Armed Police, Xi'an 710086, China; 2. Ministry of Education Key Lab. of Computer Network and Information Security, Xidian Univ., Xi'an 710071, China)

Abstract: Based on invalid probability of storage nodes, a model is provided to evaluate the reliability of a given distributed information storage system. In this model, the reliability of a storage system is affected by 4 factors: time, invalidate probability of storage nodes, data distribution algorithm and storage policy. Such a model could partly solve the reliability problem in the system designing phase, and make data service more reliable.

Key Words: access structure; failure probability; distributed storage

分布式存储是保存大量数据的常用方法. 与集中式存储相比，将数据分散到若干个相互独立的存储节点中保存可以提高数据服务的安全性和可靠性，因此分布式的存储已成为存储系统设计的主流. 存储节点通常包括 CPU、总线设备、磁盘和其他一些部件，数据则由这些节点共同保存. 现有的分布式存储架构如 SAN (Storage Area Network)和 NAS(Network Attached Storage)等都由许多个存储节点组成. 若一个存储系统中的节点的软硬件组成完全相同，则称其为同构系统，否则称为异构系统. 在实际应用中，绝大部分系统都是异构系统. 这是因为随着技术的发展，不仅仅在广域范围内使用的设备非常复杂，而且在同一个 SAN 内部都会存在使用不同操作系统的服务器和来自不同厂商的存储设备. 在一个存储系统中，通常会包括不同类型的服务器，不同的存储设备，不同的接入方法以及不同的接入协议(iSCSI, ESCON, FC, SSA, Infiniband 等)，由此导致了各个存储节点具有不同的存储容量、读写速度、故障概率以及数据传输速度. 当系统处于不可信环境中时，各个节点受到攻击的概率以及被攻破的难度也将有所不同. 因此异构存储系统将成为一个必然的发展趋势. 研究异构存储系统的性能，并据此设计存储策略具有很高的现实意义和应用背景.

可靠性是系统性能中的重要组成部分. 对于存储系统而言，数据服务的可靠性是可靠性研究的核心. 存储系统的可靠性描述了系统能有效地提供数据服务的能力，用系统能正常提供服务的概率表示.

笔者提出一种度量异构存储系统可靠性的概率模型，存储系统的可靠性将直接影响到存储服务的效率和数据的可用性. 通过对现有系统的可靠性进行预测，可以帮助存储系统的设计者和用户制定高可靠的数据

收稿日期：2008-02-23

基金项目：国家自然科学基金资助(60503012，60842006，60743005)

作者简介：张 薇(1976-)，女，武警工程学院讲师，博士，E-mail：zhaangweei@yeah.net.

存储策略以及高可靠的数据调用策略.

1　相关工作

冗余是提高存储系统可靠性的主要方法,冗余可以通过备份、数据分割、门限方案、纠错编码和纠删编码等技术来实现.冗余的存储结构可以保证部分服务器失效时数据服务仍可正常进行.

文献[1-2]分别从不同的角度分析了使用纠删编码对文件进行编码存储所能达到的数据可用性.

文献[1]中用可用概率来刻画数据对象的可用性.当系统宕机率未知时,使用纠删编码存储的文件数据的可用性为

$$P=\sum_{i=m}^{n}\binom{n}{i}p^i(1-p)^{n-i}\quad,\tag{1}$$

其中 P 为文件对象可用概率,p 为服务器节点的可用概率,n 为文件对象编码后所有的文件份额数目,m 为重构文件对象所需要的份额数目.假设系统中的机器可用概率平均为 0.7,对于使用($m=16$,$n=32$)的纠删编码系统,根据式(1)进行计算,得到的可用概率$P=0.994\,758\,998$,而对于完全备份系统来说,可用性仅为 0.91.由此看出,在冗余量相同的情况下,采用纠删编码与备份这两种方式所能达到的系统可用性相差很大.

文献[2]中讨论了假定系统中一部分服务器节点发生故障的情况下,系统中的文件所具有的访问可用性.文献[3]中的工作侧重于解决节点的可靠性评估问题,即获取节点的状态,看其是否可靠,是否已被损毁或受到了攻击,从而使数据的调用服务具有较高的可靠性.文献[4]中利用平均首次数据丢失时间(Meantime to first data loss,缩写 MTTDL)来衡量系统的可靠性,并比较了在使用条带、复制和纠删编码的存储系统中,第一次数据丢失的平均时间.

在调用数据时的节点选择方面,现有的研究工作[5-7]都集中于提高系统匿名性以及防止 P2P 系统受到攻击.其他研究如 Scan[8],集中于设计副本替代算法,以满足客户延迟和服务器负载限制.文献[3]中构造了 RDSS,一种基于 RAN(Resource Area Network)的存储系统,作者阐述了节点选择的重要性,并设计了 NRS(Node Ranking System)来管理节点,NRS 不仅可以满足负载均衡及稳定性需求,还能避免用户与有恶意的服务器通信.NRS 的目标是提高 RDSS 的可靠性,提高性能,减少与有恶意的服务器通信的开销.

与以上工作相比,笔者的工作侧重于整个分布式存储系统的可靠性评估,侧重于解决给定系统的可靠性描述以及计算方法,再以可靠性研究的结果来影响存储策略的制定,包括数据分离算法与存取结构的选择以及节点选取策略.在此基础上,制定更为可靠的数据调用策略.

2　分布式存储系统的可靠性

2.1　存储服务与存取结构

存储服务包括存储策略和数据调用策略以及必要的安全机制.存储策略主要解决保存数据时的节点选择问题,而调用策略主要解决调用数据时的内容搜索和节点选择问题.另外存储服务还包括数据更新策略、用户管理、节点管理等等.

对于同构存储系统,由于各节点完全相同,保存和调用数据时可以随机选择节点.而在异构存储系统中,不同的节点具有不同的可靠性及性能,为了提高数据服务的安全性、可靠性及效率,在保存和调用数据时必须选择适当的存储及调用策略.

在分布式存储系统中,数据通过某种算法分为若干个份额,并分散到多个存储节点保存,常见的分离算法有备份、数据分割或条带技术、门限方案、纠删编码等.份额与物理存储节点间的对应关系称为存储策略.

假设要将数据 D 保存到 n 个节点,每个节点有自己的 ID 号,其中保存的数据称为一个份额.

定义 1　设 G_1 为原始数据集合,G_2 为份额集合,则数据分离算法定义为一个映射 P.

$$P:G_1\rightarrow G_2\times G_2\times\cdots\times G_2\quad,$$

$$P(D)=(d_1,d_2,\cdots,d_n)\quad,$$

其中 $D \in G_1, d_i \in G_2$, $1 \leqslant i \leqslant n$.

相应的数据恢复算法定义为 $C: G_2 \times G_2 \times \cdots \times G_2 \to G_1$,

$$C(d_1, d_2, \cdots, d_k) = D \quad ,$$

其中 n 和 k 为自然数且 $n \geqslant k$.

在分布式存储系统中,存取结构定义为存储节点集合的子集,从该子集中的节点保存的份额集合可以恢复出原始数据.

定义 2 设 S 为存储节点集合,$\Gamma \in 2^S$,设 $\Gamma = \{s_1, s_2, \cdots, s_k\}$,$\Gamma$ 中各节点保存的数据为 $d_1, d_2, \cdots, d_k$,则 Γ 为一个存取结构当且仅当从 $d_1, d_2, \cdots, d_k$ 中可以恢复出原始数据 D.

存取结构是指可以恢复数据的存储节点集合. 由定义,任意一个存取结构与任意一个节点的并集仍为一个存取结构,在此引入最小存取结构的概念.

定义 3 最小存取结构是这样的一个存取结构,如果从其中任意去掉一个节点,则剩余的节点集合不再构成存取结构.

在存储系统中,一旦数据分离算法给定,便可以计算出所有的份额,同时,也确定了由哪些份额能构成存取结构. 因此,数据分离算法决定了每个份额参与的最小存取结构的数量,或者说,决定了每个份额的"重要性". 笔者不考虑具体的算法,而只根据份额的"重要性"来确定系统可靠性,从而为存储策略的制定提供依据.

若存储系统采用完全复制技术,则每个节点都是一个最小存取结构. 如果采用秘密共享,则每个节点中只保存一个份额. 设系统中共有 n 个份额 $s_1, s_2, \cdots, s_n$,这些份额可构成 M 个最小存取结构(AS),假设 n 个份额参与的最小存取结构数目分别为 $m_1, m_2, \cdots, m_n$,参与了 j 个 AS 的份额数为 l_j,$1 \leqslant j \leqslant M$, 则有如下一些基本关系成立:

(1) $\sum_{i=1}^{n} m_i = kM$,其中 k 为一个 AS 中包含的节点数(只适用于 AS 中节点数目相同的情形);

(2) $\sum_{j=1}^{M} l_j = n$;

(3) $\sum_{j=1}^{M} j l_j = \sum_{i=1}^{n} m_i$.

2.2 存储系统的可靠性

可靠性理论的基本问题在于预测系统什么时候最终失效. 组件失效时间的不确定性用概率密度函数 $f(t)$ 来描述[9],而分布函数 $F(t)$ 是时刻 0 到时刻 t 失效的概率,记为

$$F(t) = \int_0^t f(l)\,\mathrm{d}l \quad .$$

可靠性也可以用平均无失效运行时间和平均修复时间来度量[9].

存储系统的可靠性体现了系统能有效地提供数据服务的能力,可以根据存储系统的组件特性来预测整个系统的可靠性. 与软件可靠性类似,用概率密度函数描述存储节点失效时间的不确定性,由此可以构造节点失效的概率模型,进而得到整个系统的可靠性.

可靠性可以用正常数据服务的概率表示,在这里有两个基本的假设:

(1) 若一个节点出现故障,则认为其中保存的数据均不可用;

(2) 若一个节点受到攻击,则认为其中保存的数据均不可用.

在这样的假设条件下,可以认为节点在遭受攻击或者出现损毁时对数据造成的破坏是相同的,均导致该节点上保存的所有数据不可用.

存储系统的可靠性有这样几个原则:

(1) 对于同一个系统,有 i 个节点失效时的可靠性不低于有 $i+1$ 个节点失效时的可靠性,所有节点均能正常工作时,系统的可靠性最高;

(2) 可靠性是关于时间 t 的单调递减函数;

(3) 假设将数据 D 分为 n 个份额,保存到 n 个节点,并且这 n 个节点可以构成 m 个最小存取结构,则 m 越

大,数据服务也越可靠.同时,对于每个节点而言,该节点参与的存取结构越多,则对于系统可靠性的影响就越大.

根据上述分析,笔者提出两个可靠性度量模型.

2.2.1　简单概率模型

一个节点的失效概率是其损毁概率和受到攻击的概率之和.在简单概率模型中,假设这些概率是先验地测定的,并具有固定的值,且各节点之间相互独立.

设系统中共有 n 个节点,各个节点的损毁概率分别为 $p_{11},p_{12},\cdots,p_{1n}$,受到攻击的概率分别为 $p_{21},p_{22},\cdots,p_{2n}$,则各节点处于正常工作状态的概率为

$$q_1=1-(p_{11}+p_{21}),\ q_2=1-(p_{12}+p_{22}),\ \cdots,\ q_n=1-(p_{1n}+p_{2n}) \quad .$$

在同构存储系统中,各 p_{1i} 及各 p_{2i} 均相等.而在异构系统中,至少存在一对整数 i, j,使得 $p_{1i}\neq p_{1j}$ 或 $p_{2i}\neq p_{2j}$.

定义 4　具有 n 个节点的分布式系统的状态是一个 n 维向量 $\boldsymbol{\theta}=(a_1,a_2,\cdots,a_n)$,$a_i\in\{0,1\}$,其中 $a_i=0$ 表示第 i 个节点不可用,即处于损毁或被攻击状态,而 $a_i=1$ 表示第 i 个节点可以正常工作.

设 $P_{\boldsymbol{\theta}}$ 为状态 $\boldsymbol{\theta}$ 出现的概率,则

$$P_{\boldsymbol{\theta}}=\prod_{\substack{i=1\\a_i=0}}^{n}(p_{1i}+p_{2i})\prod_{\substack{i=1\\a_i=1}}^{n}q_i \quad .$$

存储系统在某一状态 $\boldsymbol{\theta}$ 下是可用的当且仅当在该状态下至少有一个可用的存取结构,此时系统为可靠状态.令概率 P_a 表示系统中至少存在一个可用存取结构的概率,即系统可靠工作的概率.则系统的可靠性可以用这个概率来度量,即

$$P_a=\sum_{\boldsymbol{\theta}}p_{\boldsymbol{\theta}} \quad ,$$

其中 $\boldsymbol{\theta}$ 为可靠状态,而整个系统失效的概率为 $1-P_a$,P_a 的取值范围在 0 到 1 之间,P_a 越大,则系统越可靠.

例 1　假设系统中共有 5 个存储节点,采用(3,5)门限方案来分离数据,则有 10 种可能的最小存取结构,分别为{1,2,3},{1,2,4},{1,2,5,},{1,3,4},{1,3,5},{1,4,5},{2,3,4},{2,3,5},{2,4,5},{3,4,5}.5 个节点构成的系统共有 $2^5=32$ 种状态,其中只有 16 个向量的重量不小于 3,故只有 16 种可靠状态.若系统是同构的,则每个节点正常工作的概率相同,设其为 0.9,而故障概率为 0.1,可以求出此时的数据可靠度 P_a 为

$$P_a=0.9^5+C_5^4\times0.9^4\times0.1+C_5^3\times0.9^3\times0.1^2=$$
$$0.59049+0.32805+0.0729=0.99144 \quad .$$

若存储系统为异构系统,则 P_a 的计算要更复杂一些.比如,节点 1 与其他节点正常工作的概率不同,节点 1 正常工作的概率为 0.8,其余同上,则此时的可靠度为

$$P_a=0.9^4\times0.8+\lfloor0.9^4\times0.2+4\times0.9^3\times0.1\times0.8\rfloor+$$
$$\lfloor6\times0.9^2\times0.8\times0.1^2+4\times0.9^3\times0.1\times0.2\rfloor=$$
$$0.52488+0.13122+0.23328+0.03888+0.05832=0.98658 \quad ,$$

可靠性明显降低了.

由此可见,在异构存储系统中,若要保持较高的可靠度,必须合理选择存储策略.若某个节点易于被攻击或受到破坏,则减少该节点参与的存取结构数量,以便提高可靠度.而这一点一般的门限方案是做不到的,必须设计其他的数据分离算法,使各份额参与的存取结构数量不同[10].

2.2.2　基于失效概率密度函数的模型

在实际应用中,节点的失效概率是随时间变化的一个不确定的量,而简单概率模型无法体现这一点,因此需要定义节点的失效概率密度函数.如果将时间考虑进来,则在时刻 t,节点 i 失效的概率是关于 t 的一个函数.对于同构系统,各节点的失效概率密度函数相同,均为 $f(t)$,而节点在时刻 T 失效的概率为从时刻 0 到 T 之间函数 $f(t)$ 的积分值.这样便满足了系统可靠性的第二个原则,即系统的可靠性随时间延长而降低.

对于异构系统,假设每个节点关于时间 t 的失效概率密度函数是独立的,并设节点 i 的失效概率密度为

$f_i(t)$,则节点 i 在时刻 T 的失效概率为$F_i(T)=\int_0^T f_i(t)\,\mathrm{d}t$.

在由 k 个节点构成的集合 A 中,所有节点在时刻 T 全部失效的概率定义为联合失效概率 $F_C(T)$,即 $F_c=F_1(T)F_2(T)\cdots F_k(T)$.该节点集合的可靠度 D_A 定义为该集合中所有节点均正常工作的概率

$$D_A=(1-F_1(T))\cdots(1-F_k(Tt))\quad .$$

设给定的存储节点集合为$\{\mathrm{node}_{i1},\mathrm{node}_{i2},\cdots,\mathrm{node}_{ik}\}$,则该集合与系统状态

$$\underset{(0\cdots0}{1}\ \underset{1}{i_1}\ \underset{0\cdots0}{}\ \underset{1}{i_2}\ \underset{0\cdots0}{}\ \underset{1}{i_k}\ \underset{0\cdots0)}{}$$

相对应.

设 T 时刻的系统状态 $\boldsymbol{\theta}=(a_1,a_2,\cdots,a_n)$,$a_i\in\{0,1\}$,则 $\boldsymbol{\theta}$ 出现的概率为

$$P_{\boldsymbol{\theta}}=\prod_{\substack{i=1\\a_i=0}}^{n}(p_{1i}+p_{2i})\prod_{\substack{i=1\\a_i=1}}^{n}q_i\quad .$$

节点 i 在时刻 T 失效的概率为$F_i(T)=\int f_i(t)\,\mathrm{d}t$,正常工作的概率为 $1-F_i(T)$,由此得到 $P_{\boldsymbol{\theta}}$ 的另外一种表达式为

$$P_{\boldsymbol{\theta}}=\prod_{a_i=0}F_i(T)\prod_{a_i=1}(1-F_i(T))\quad .$$

当系统处于某个状态 $\boldsymbol{\theta}$ 时,如果从可用节点集合中去掉节点集合 Γ,其余的可用节点恰好构成一个最小存取结构,则称 Γ 为状态 $\boldsymbol{\theta}$ 可以容忍的一个失效节点集合.由于最小存取结构不是惟一的,因此,一个状态可以容忍多个失效节点集合.对于一个任意给定的节点集合,这个集合的可靠度取决于该集合可以容忍多少个节点失效以及这些节点的失效概率.事实上,一种状态对应着一个可用节点集合.因此状态的可靠度也取决于在该状态下可以容忍的失效节点数.

设状态 $\boldsymbol{\theta}$ 可以容忍的失效节点集合为$\Gamma_{\boldsymbol{\theta}}^1,\cdots,\Gamma_{\boldsymbol{\theta}}^r$,共 r 个,而这 r 个节点集合失效的概率是不同的.当且仅当这些集合中的每个节点均失效时,系统达到临界状态,此时系统中的可用节点恰好构成一个最小存取结构,再增加一个失效节点便会导致数据不可用.

设系统在状态 $\boldsymbol{\theta}$ 下达到临界状态的概率为 $P_{\boldsymbol{\theta}}^C$,令$\Gamma=\Gamma_{\boldsymbol{\theta}}^1\cup\cdots\cup\Gamma_{\boldsymbol{\theta}}^r=\{s_1,s_2,\cdots,s_w\}$,则 $P_{\boldsymbol{\theta}}^C=\prod_{i=s_1}^{s_w}F_i(T)$.显然,$P_{\boldsymbol{\theta}}^C$ 越大,系统越容易到达临界状态,因而也越不可靠.据此给出可靠性的第二个度量标准:

$$P_{\mathrm{s}}=\sum_{\boldsymbol{\theta}}P_{\boldsymbol{\theta}}(1-P_{\boldsymbol{\theta}}^C)\quad ,$$

式中 P_{s} 为系统的可靠度.

由以上分析知,可靠性函数是一个关于时间 T 的函数,它与 3 个因素有关:时间,最小存取结构,节点失效概率密度.而最小存取结构又由数据分离算法和存储策略决定,因此可以选择适当的数据分离算法和存储策略来影响可靠度函数的取值.从而在保存数据时,根据先验测得的节点失效概率来选择数据分离算法及保存策略,可以使系统可靠性达到最大值.

2.3 分布式存储系统体系结构

上文构造的可靠性模型在实际存储系统中可以起到可靠性预测的作用,为用户制定存储策略提供信息.用户在制定存储策略之前首先利用模型对各个节点的可靠性进行分析,选择那些可靠性较高的节点提供数据服务.通过可靠性分析,可以调整存储策略,以最大限度地实现服务的持续性.

图 1 中给出了一种分布式存储系统体系结构,可以作为可靠性模型的具体应用环境,系统中包含存储节点、用户、访问控制系统,数据读写协议、访问控制、可靠性分析、入侵检测及故障检测等组件.

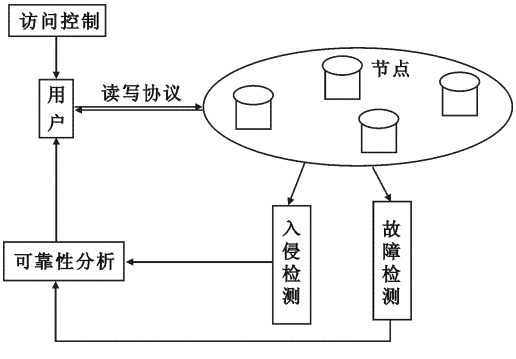

图 1 分布式存储系统体系结构

假设系统中任何一个存储节点都有可能出现故障或者遭到攻击,系统的当前状态由入侵检测和故障检测部件来判断,入侵检测及响应模块用于检测系统中被入侵的节点,故障检测用于检测节点是否有故障.检测的结果送入可靠性分析模块,用于计算节点的可靠程度.除了正常工作状态,其他情形一概被视为失效,节点的可靠性可以用失效概率来表示,概率越大,则失效的可能性越高,在选择时应尽量避开该节点.这三者共同为节点选择提供依据,从而制定更可靠的存储策略,并将可靠性问题部分地解决于系统设计和初始化阶段.

3 总　　结

通过对存储系统整体可靠性的研究,分析了影响可靠性的因素,包括时间、节点失效概率密度函数、分离算法及存储策略等,提出了存储系统的可靠性模型,在具体存储系统中,该模型与入侵检测、故障检测等部件相结合,可以在系统设计及初始化阶段解决可靠性问题,使数据服务在发生部分故障时能持续进行.

参考文献:

[1] Siewiorek D P, Swarz R S. Reliable Computer Systems: Design and Evaluation[M]. Burlington: Digital Press, 1992: 35-47.

[2] Weatherspoon H, Kubiatowicz J. Erasure Coding vs Replication: a Quantitative Comparison[C]//The 1st Workshop on Peer-to-Peer Systems. Cambridge: Springer, 2002: 328-337.

[3] Li Xiaodong, Liu Chang. Towards a Reliable and Efficient Distributed Storage System[C]//Proc of the 38th Annual Hawaii International Conference on System Sciences(HICSS'05). Big Island: IEEE, 2005: 301-311.

[4] Svend F, Arif M. A Decentralized Algorithm for Erasure-Coded Virtual Disks[C]//Proceedings of Dependable Systems and Networks. Florence: IEEE, 2004:125-134.

[5] Cornelli F, Dimiani E. Choosing Reputable Servernts in a P2P Network[C]//Proc of the 11th International World Wide Web Conference. Hawaii: Springer, 2002: 137-145.

[6] Ernesto D, Sabrina D C, Stefano P, et al. A Reputation-based Approach for Choosing Reliable Resources in Peer-to-peer Networks[C]//Proc of the 9th ACM Conference on Computer and Communications Security. Washington: ACM Portal Press, 2002: 207-216.

[7] Dingledine R, Freedman M J, Molnar D. The Free Heaven Project, Distributed Anonymous Storage Service[C]//Proc of the Workshop on Design Issues in Anonymity and Unobservability. California: Springer, 2000: 308-319.

[8] Chen Y, Katz R H, Kubiatowicz J D. SCAN: a Dynamic Scalable and Efficient Content Distribution Network[C]//Proceeding of the 1st International Conference on Pervasive Computing. Zurich: Springer, 2002: 282-296.

[9] Fenton N E, Pfleeger S L 著. 软件度量[M]. 杨海燕, 赵巍, 译. 第一版. 北京: 机械工业出版社, 2004: 1-10.

[10] 张薇, 马建峰. LPCA——分布式存储中的数据分离方法[J]. 系统工程与电子技术, 2007, 29(3): 19-24.
Zhang Wei, Ma Jianfeng. LPCA—Data Distribution Algorithm in Distributed Storage[J]. Systems Engineering and Electronics, 2007, 29(3): 19-24.

[11] 樊鹤红, 孙小菡. 一种通用的网络可靠性仿真模型[J]. 西安电子科技大学学报, 2007, 34 (Sup): 68-71.
Fan Hehong, Sun Xiaohan. A General Network Reliability Simulation Model[J]. Journal of Xidian University, 2007, 34 (Sup): 68-71.

(编辑: 郭　华)