

基于框图法的网络存储系统可靠性分析

张 英¹ 张 益² 王冀鲁²

(中国科学院计算技术研究所 北京 100190)¹

(中国矿业大学(北京)机电与信息工程学院 北京 100083)²

摘 要 随着信息的快速增长和对数据安全要求的提高,网络存储系统的可靠性直接影响信息系统性能。重点针对网络存储系统中的设备失效问题,在分析典型网络存储系统可靠性模型的基础上,采用框图法理论提出了一种基于混联结构的网络存储系统的可靠性分析与评价方法,包括基于属性节点的系统可靠性框图描述方法、负载关联的 M 取值方法等。通过典型混联结构的 FC-SAN 系统实验,验证了该方法的有效性。该方法相对于测试等其它方法具有简单、省时、费用低等特点,因此更有利于应用于网络存储系统的方案和设备优化。

关键词 网络存储系统,可靠性评估,框图法

中图法分类号 TP393 文献标识码 A

Reliability Evaluation of the Network Storage System Based on Block Diagrams

ZHANG Ying¹ ZHANG Yi² WANG Ji-lu²

(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)¹

(School of Mechanical Electronic & Information Engineering, China University of Mining & Technology (Beijing), Beijing 100083, China)²

Abstract With the rapid growth of information and improving data security requirements, the reliability of network storage system directly affects the information system performances. This paper focused on network storage system's equipment failure. Based on the reliability analysis about the typical model of network storage system and the use of reliability block diagrams theory, the paper proposed a hybrid form of network storage system reliability analysis and evaluation methods, including system reliability block diagram analysis methods based on nodes attribute, M value methods with the business load. And the hybrid structure of FC-SAN systems experiment was established to verify the effectiveness of the method. As a result of this method compared to other methods of testing with a simple, time-saving, low cost, it is more conducive to application of network storage system to optimize the scheme and equipment.

Keywords Network storage system, Reliability evaluation, Reliability block diagrams

1 引言

在数字化和网络互联时代,随着网络业务系统对数据依赖性的增强,数据的集中管理、综合分析、容灾备份等处理要求日益提高;随着信息的快速增长和对安全要求的增高,人们对存储系统的需求量不断增大,网络存储已成为各类信息系统的重要组成部分,从而对网络存储系统的性能和可靠性提出了越来越高的要求。

数据存储系统的发展经历了 3 个阶段,即直接外挂存储(DAS)、网络附加存储(NAS)和存储域网络(SAN)^[1]。随着数据存储系统性能的不提高,其系统的可靠性直接影响信息系统性能。如何有效地规划和配置高可靠性的网络存储系统,不仅可以简化异构存储管理的复杂性,更可以高效、充分地利用存储设备。因此,研究网络存储系统的可靠性分析方法,可为配置高可靠性的存储系统提供理论依据。

可靠性技术经历了 50 多年的应用和发展并逐步走向成熟。各国对可靠性研究都非常重视,该技术已大量应用在军

事、电子信息、能源等领域。我国近十几年来在可靠性方面的研究不断兴起。可靠性框图法(RBD-RELIABILITY BLOCK DIAGRAMS)^[2]主要是一种应用于设备可靠性分析的方法,它大多应用于电子产品、机械部件、医疗器械等设备方面的可靠性评价,以及系统安全领域。这种方法通过方框图来表示系统的可靠性结构,是一种简单的、描述功能结构和故障单元对系统影响的图形表现方法。

本文重点针对网络存储系统中的设备失效问题,在分析典型网络存储系统可靠性模型的基础上,采用框图法理论,提出一种基于混联结构的网络存储系统的可靠性分析与评价方法,包括基于属性节点的系统可靠性框图描述方法、面向负载关联的 M 取值方法等。通过对典型混联结构的 FC-SAN 系统实验,验证了本方法的有效性。

2 存储系统可靠性评价指标

系统的可靠性:在给定的时间内计算机系统能实施应有功能的能力。系统的可靠性常用平均无故障时间(MTBF)来

到稿日期:2009-07-08 返修日期:2009-09-24 本文受科技部技术创新基金项目(BT2008-22)资助。

张 英(1951—),女,研究员,主要研究方向为计算机网络及性能评价,E-mail:ccss@ict.ac.cn;张 益(1983—),男,硕士,主要研究方向为计算机网络及应用;王冀鲁(1973—),男,博士生,主要研究方向为计算机网络及应用。

度量^[3],即系统平均能够正常运行多长时间才发生一次故障。系统的可靠性越高,平均无故障时间越长。

可靠性:计算机系统的可靠度是指从它开始运行($t=0$)到某时刻 t 这段时间内能正常运行的概率,用 $R(t)$ 表示。

失效率:是指单位时间内失效的元件数与元件总数的比例,以 λ 表示。当 λ 为常数时,可靠度与失效率的关系为

$$R(t) = e^{-\lambda t}$$

MTBF:两次故障之间系统能正常工作的时间的平均值称为平均无故障时间(MTBF)。

$$MTBF = 1/\lambda$$

3 框图法基本结构

可靠性方框图(RBD-RELIABILITY BLOCK DIAGRAMS)通常用于表示系统的可靠性结构。RBD是一种表示所有可能的功能结构以及故障的单元对系统功能影响的图形方法。它主要是用于计算设备的可靠度。通常包括以下3种基本结构^[4,5]。

1)串联系统:假设一个系统由 N 个子系统组成,当且仅当所有的子系统都能正常工作时,系统才能正常工作,这种系统称为串联系统。

设各子系统的可靠性为 $R_1, R_2, \dots, R_n$,则整个串联系统的可靠性为

$$R = R_1 \times R_2 \times \dots \times R_n$$

设各子系统的失效率为 $\lambda_1, \lambda_2, \dots, \lambda_n$,则整个串联系统的失效率为

$$\lambda = \lambda_1 + \lambda_2 + \dots + \lambda_n$$

2)并联系统:假如一个系统由 N 个子系统组成,只要有一个子系统正常工作,系统就能正常工作,这样的系统称为并联系统。

设各子系统的可靠性为 $R_1, R_2, \dots, R_n$,则整个并联系统的可靠性为

$$R = 1 - (1 - R_1)(1 - R_2) \dots (1 - R_n)$$

设各子系统的失效率均为 λ ,则整个并联系统的失效率为

$$\mu = \frac{1}{\frac{1}{\lambda} \sum_{j=1}^n \frac{1}{j}}$$

3)N模冗余系统:N模冗余系统由 N 个($N=2n+1$)相同的子系统和一个表决器组成,表决器把 N 个子系统中占多数的相同结果的输出作为系统的输出。

设各子系统的可靠性均为 R_0 ,则整个N模冗余系统的可靠性为

$$R = \sum_{i=n+1}^N \binom{j}{N} \cdot R_0^i (1 - R_0)^{N-i}$$

4 典型存储系统的可靠性分析

4.1 RAID 可靠性分析

RAID是利用若干硬盘驱动器加上控制器按一定的组合方式组成的一个大容量、快速响应和高可靠的存储子系统。硬盘驱动器的不同组合方式形成了不同级别的RAID阵列^[6]。

1)RAID0

RAID0方式的数据流分别通过不同的驱动器进行传输,

如果其中的一个硬盘驱动器发生故障,整个硬盘数据就被破坏。显然,系统的可靠性模型符合串联模型,其可靠性值是每一个单独硬盘驱动器可靠性的乘积。设有 n 块磁盘,每块可靠性为 r ,则RAID0的可靠性 $R=r^n$ 。

2)RAID1

RAID1方式是通过磁盘镜像将所有数据复制到另一个硬盘驱动器上,所以RAID1的可靠性模型为先串联再并联。

3)RAID3

在RAID3方式中,将奇偶校验信息存储于独立的硬盘驱动器,而数据被存储到其它的硬盘驱动器上。这种方式至少需要3块硬盘驱动器。在由 n 块硬盘组成的RAID3磁盘阵列中,如果有一个硬盘驱动器出现故障,数据不会丢失;如果奇偶校验硬盘驱动器出现故障,存储的数据不会丢失,但冗余将丢失。RAID3符合N模冗余系统模型。

4)RAID5

在RAID5方式中,奇偶校验信息与数据被保存到所有的硬盘驱动器上,从而消除了单个奇偶校验硬盘驱动器的瓶颈问题。RAID5磁盘阵列的性能比RAID3有所提高,但仍然需要至少3块硬盘驱动器。在RAID5阵列中,当一个硬盘驱动器发生错误时,所有的数据仍然是可用的,丢失的数据将从剩余的硬盘驱动器和奇偶校验信息中重新计算得到。但当2个以上的硬盘驱动器同时出现故障,则所有数据将会丢失。N块磁盘组成的RAID5系统可靠性模型符合N模冗余系统模型。

4.2 DAS结构的可靠性分析

DAS^[7]是最简单的一种数据保护方案,在大多数情况下,这种备份大多是采用服务器上自带的磁带机或备份硬盘,而备份操作往往也是通过手工操作的方式进行的。DAS系统逻辑模型符合串联结构,如图1所示。

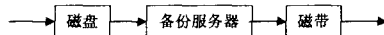


图1 DAS系统逻辑图

4.3 NAS结构可靠性分析

在NAS结构中,客户的请求通过网络交换机到达服务器,服务器将根据任务要求对存储设备发起数据存取请求,每次数据存储完成后均由交换机向服务器返回应答,再由服务器向客户返回结果。从数据流程看,该系统的逻辑模型由NAS设备、LAN、server三者串联组成,如图2所示。



图2 NAS系统逻辑图

4.4 SAN结构可靠性分析

在SAN^[8]结构中,客户的请求通过局域网经由与服务器相连的交换机端口到达服务器,由服务器处理后,服务器将根据任务要求通过存储域网对存储设备发起若干次数据存取请求。服务器处理完成后,将结果从与服务器相连的交换机端口返回给客户端。SAN系统的逻辑模型如图3所示。



图3 SAN系统逻辑图

5 基于混联结构的网络存储系统可靠性分析

框图法主要用于计算设备的可靠度,一般采用上述提到

的3种基本结构进行计算,主要运用于电子产品、机械部件、医疗器械等设备可靠度预计方面。另外,在系统安全领域也有相关运用。

在网络存储系统中,其设备的互联是一种串并联混杂的形式。本文通过分析现有典型网络存储系统的可靠性模型,给出基于混联结构的网络存储系统的可靠性分析与评价方法,包括基于属性节点的系统可靠性框图描述方法、面向负载关联的M取值方法等。其典型混联结构的FC-SAN系统的逻辑模型如图4所示。

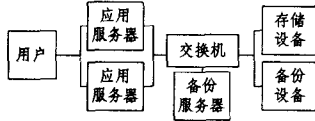


图4 混联方式的网络存储系统拓扑图

该系统模型主要有以下特点:

- 1)典型性。串并联加备份是当前网络存储系统应用的典型结构;
- 2)普遍性。包括了可靠性框图模型中的串联和并联两类连接方式;
- 3)简洁性。简化当前复杂的网络存储结构,便于试验和分析对象模型的可靠性;
- 4)可扩展性。基于典型的模型结构,可以方便地向大型复杂的网络环境中扩展。

5.1 面向负载关联的M取值方法

在可靠性分析中,常用指数分布、正态分布、对数正态分布和威布尔分布等作为寿命分布^[9]。在这些常用的分布函数中,威布尔分布被广泛应用于研究机械、化工、电气、电子、材料的失效分析。

网络存储系统中设备的失效性用单元设备的可靠度表示,采用双参数的威布尔分布作为寿命分布函数,其函数定义如下^[10]:

$$R(t) = \exp\left\{-\left(\frac{t}{\eta}\right)^m\right\}, t \geq 0, m > 0, \eta > 0$$

式中, m 和 η 分别是分布的形状(斜率值)和尺度(特征寿命)参数。 η (特征寿命)定义为单元期望失效(MTBF)的63.2%时的寿命。

m 值表现失效率变化的参数,它的取值不同,概率密度函数的曲线形状也不同。在进行一般设备的失效性分析中,通常 m 值是按照设备失效的不同时期进行取值的。如早期故障($m < 1.0$)、随机失效($m = 1.0$)、早期损耗($1.0 < m < 4.0$)、快速损耗($m > 4.0$)等。

在网络存储系统的设备失效性分析中,除考虑使用寿命的影响外,更重要的应考虑负载对设备失效性的影响。为了解决网络存储系统中业务压力对故障率的影响问题,以便找到业务量变化与斜率值 m 的关系,我们采用将不同的业务量与 m 值成正比对应的关系,并依据设备的性能等级,将业务量的类别(每秒并发I/O数、吞吐量Mbps、平均每秒并发请求数、并发用户数)分为5个等级,依据业务量从低到高的顺序分别对应的 m 值为0.5,1,2,3.4和5。在此基础上分别对7个厂家的8类存储设备进行了测试实验,实验结果如图5所示。

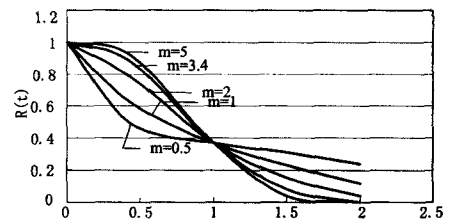


图5 设备可靠度与 m 值随时间的变化关系图

通过以上实验结果可见,当运行时间小于特征寿命($t/\eta < 1$)时, m 值越大,可靠性越高。这是因为在特征寿命期间内,由于设备处于使用初期,在此期间故障率大致处于稳定状态,趋于定值,因此与业务量关系不大。

当运行时间大于特征寿命($t/\eta > 1$)时,由于设备随着使用时间的延长,因磨损、疲劳、加剧而使设备故障率逐渐上升,说明设备已到了使用寿命期,所以随着业务量的增大,设备负荷增大, m 值越大,可靠性越低。

根据实测结果分析,同时考虑各时期业务流量特点,按照业务量与设备性能参数的比值,存储系统的 m 取值可在0.5~4之间选定,如表1所列。

表1 存储设备 m 值分级表

性能分级		业务量分级	M取值
存储设备	高	100%	4
		80%	3
		50%	2
		10%	1
		5%	0.5
	中	100%	4
		80%	3
		50%	2
		10%	1
		5%	0.5
	低	100%	4
		80%	3
		50%	2
		10%	1
		5%	0.5

5.2 基于属性节点的系统可靠性框图描述方法

在采用威布尔分布计算设备可靠度的基础上,系统级的可靠度是各单元可靠度的组合。该系统框图通常都有一个起点和一个终点,如图6所示。其中至少要有一条可以从起点到达终点的路径,且该路径没有经过一个故障单元,系统才是正常的。

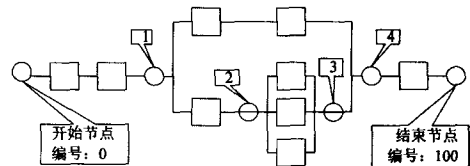


图6 系统可靠度算法示例图

由于采用串并混联方式的可靠性框图来表示网络存储系统的组成单元 $D(i, j)$,在框图描述中首先需要确定每一单元的串并联属性以及单元之间的关系,以利于系统级可靠性的计算。为此,引入了属性节点 $J(x, y)$ 、左接口、右接口概念来解决获取系统各单元串并联属性的问题。

描述定义:

- 1)组成单元 $D(i, j)$ 和属性节点 $J(x, y)$ 包括左接口、右接口,接口表示连接状态。

2) $D(i, j)$ 的左右接口分别只能有一种连接状态, $J(x, y)$ 可以有多种连接状态。

3) $J(x_1, x_2)$ 需成对出现, 两节点间构成子系统, 子系统可以分别由串联或并联系统组成。

4) 当组成单元 D_i 和 D_j 之间存在并联关系时, 需增加属性节点 J 。

5) 元件归一处理, 将计算后的多个组成单元合成, 并删除多余节点。

根据以上定义, 系统可靠度表示为

$$R = R_s \cdot R_p$$

式中, R_s 表示串联子系统可靠度, R_p 表示并联子系统可靠度。

各子系统可靠度为

$$R_{s1} = R_{11} \times R_{12} \times \dots \times R_{1n}$$

$$R_{p1} = 1 - (1 - R_{11})(1 - R_{12}) \dots (1 - R_{1n}), \dots$$

5.3 实验及结果

5.3.1 实验系统结构

依据以上方法, 建立了网络存储系统的实验模型, 该实验模型给出了具有冗余和串并联混合的 FC-SAN 系统。在分析中, 可根据实验模型中的单元不可靠度值定义可靠性脆弱点; 可计算出在业务量不变的情况下系统可靠性随时间的变化曲线图; 并可给出当业务持续时间相同时系统可靠性与 m 值变化的关系。

我们通过实验模型进行了模拟实验, 实验系统拓扑如图 7 所示。

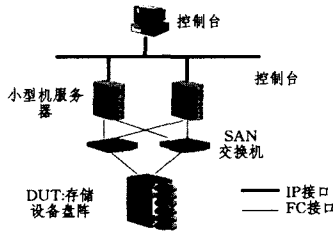


图 7 模拟系统拓扑图

实验环境主要由以下几部分组成。控制台: 1 台 DL380 (Windows2000 系统); 小型机服务器: 8 台 DL380 (Linux AS 4.4); SAN 交换机: HP128 口光纤交换机; 存储设备: HP EVA8100。

5.3.2 关键业务描述

当今的数据处理大致可以描述为两大类, 即联机事务处理 OLTP (on-line transaction processing)、联机分析处理 OLAP (On-Line Analytical Processing)。OLTP 是传统的关系型数据库的主要应用, 主要是基本的、日常的事务处理, 例如银行交易。OLAP 是数据仓库系统的主要应用, 支持复杂的分析操作, 侧重决策支持, 并且提供直观易懂的查询结果。

在建立的实验模型中, 分别对这两种业务进行了配置, 如表 2 和表 3 所列。

表 2 OLTP 业务模型

模型名称	IO Profile	块大小	比例
OLTP	随机读	8k	20%
	随机读	4k	45%
	随机写	8k	15%
	顺序读	64k	10%
	顺序写	64k	10%

表 3 OLAP 业务模型

模型编号	IO Profile	块大小	比例
OLAP	随机读	4k	0%
	随机读	4k	15%
	随机写	4k	5%
	顺序读	64k	70%
	顺序写	64k	10%

5.3.3 可靠性框图模型

从实验系统拓扑图抽象出图 8 所示的 FC-SAN 系统可靠性模型。

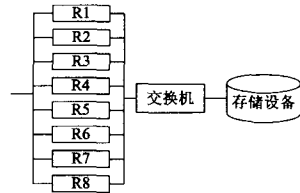


图 8 FC-SAN 可靠性模型

以上服务器 $R1 \sim R8$ 属于并联子系统, 所以 $R1 \sim R8$ 设备的可靠度为

$$R_s = 1 - (1 - R_1)(1 - R_2) \dots (1 - R_8)$$

该并联子系统和交换机、存储设备构成串联子系统, 所以依据串联系统可靠度计算公式, 则该网络存储系统的可靠度为

$$R = R_s \times r_w \times r_d$$

式中, r_w 为交换机可靠度, r_d 为存储设备可靠度。

实验环境参数: 小型机服务器 DL380 的 MTBF 为 675000h; SAN 交换机的 MTBF 为 10000h; 存储设备盘阵 HP EVA8100 的 MTBF 为 100000h, 该系统可靠度计算结果如表 4 所列。

表 4 系统可靠度计算结果

时间(h)	200	400	600	800	1000
可靠度	0.975924	0.952428	0.929498	0.907119	0.885280

当业务量负载维持在设备性能参数的 100% 时, 即 m 值为 4 时, 实验结果如图 9 所示。

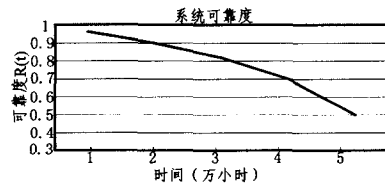


图 9 系统可靠度随时间变化图

当运行时间大于特征寿命 ($t/\eta = 1.5$) 时, 业务量与性能参数比值在 5%~100% 变化, 对应 m 值为 0.5, 1, 2, 3, 4, 此时可靠度随 m 值的变化如图 10 所示。

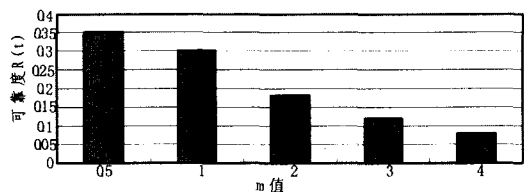


图 10 系统可靠度随 m 值变化图

结束语 本文针对网络存储系统的可靠性评价问题开展

(下转第 195 页)

组/s 情况下的平均访问时间图。在这两个图中,随着窗口的增大,每种结构的访问时间都有所增大,带分支链表的梯队列具有最好的时间性能。图 6 和图 7 与图 4 和图 5 相比,访问增幅较小,这是因为当流速和窗口增大时,join 的增长速度比 distinct 差一倍左右,所以图 4 与图 5 的平均访问时间增长速度比图 6 和图 7 大得多。

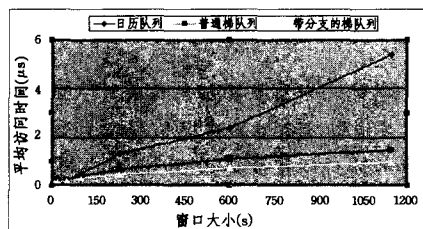


图 6 Query 2 在流入速度 5 元组/s 的平均访问时间

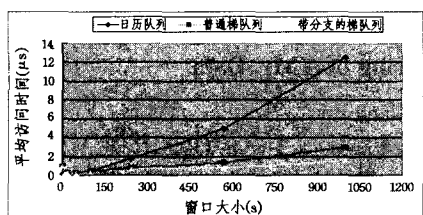


图 7 Query 2 在速度 100 元组/s 的平均访问时间

结束语 随着信息技术的不断发展,数据流在各个应用领域涌现,流入数据流查询越来越受到人们的关注。数据流上连续查询结果随着数据的不断流入流出也会不断变化。如果有效地维护数据流上连续查询的结果,很大程度上也就决定了连续查询的效率。当前鲜有研究涉及到数据流上连续查询结果的维护,特别是非单调连续查询结果的维护。

通过对连续查询结果的分析,本文设计了一种带分支链表的梯队列来维护连续查询结果。这种带分支链表的梯队列能适应具有各种不同截止期的数据的维护,而且能较大幅度地降低数据平均访问时间。实验结果表明,带分支链表的梯队列能显著提高数据流的连续查询效率。

参考文献

- [1] Babcock B, Babu S, Datar M, et al. Models and Issues in Data Streams[C]//ACM Symposium on Principles of Database Systems(PODS). 2002,5:1-16
- [2] Golab L, Ozsu M T. Issues in Data Stream Management[J]. SIGMOD Record 32,2003;5-14

(上接第 105 页)

研究,包括网络存储系统可靠性评价指标的定义,在对典型网络存储系统分析的基础上,采用框图法理论,提出基于混联结构的网络存储系统可靠性分析方法,即一种与业务负载关联的系统可靠性计算方法。该方法相对于测试等其它方法具有简单、省时、费用低等特点,因此更有利于应用于网络存储系统的方案和设备优化。

参考文献

- [1] 郑伟民,舒继武.下一代分布式智能网络存储系统的发展趋势[J].世界电信,2004,8:16-29
- [2] 曾照洋,任占勇.基于多状态的系统可靠性预计[J].Aeronautic Standardization & Quality,2008,2:43-47
- [3] 唐应辉.计算机系统的可靠性研究[D].成都:电子科技大学,

- [3] Golab L, Ozsu M T. Processing sliding window multijoins in continuous queries over data streams[C]//Proceedings of 29th International Conference on Very Large Data Bases (VLDB). 2003;500-511
- [4] Hammad M, Aref W, Franklin M, et al. Efficient execution of sliding window queries over data streams[R]. Purdue University,2003
- [5] Law Y N, Wang H, Zaniolo C. Query languages and data models for database sequences and data streams[C]//Proceedings of 30th International Conference on Very Large Data Bases (VLDB). 2004;492-503
- [6] Terry D, Goldberg D, Nichols D, et al. Continuous queries over append-only databases[C]//Proceedings of ACM-SIGMOD 1992 International Conference on Management of Data (SIGMOD 1992). 1992;321-330
- [7] Krämer J, Seeger B. A temporal foundation for continuous queries over data streams[C]//Proceedings of 11th International Conference on Management of Data (COMAD 2005). 2005;70-82
- [8] Hammad M, Mokbel M, Ali M, et al. Nile: a query processing engine for data streams[C]//Proceedings of the 20th International Conference on Data Engineering (ICDE 2004). 2004;851
- [9] Golab L, Ozsu M T. Update-Pattern-aware Modeling and Processing of Continuous Queries [C]//Proceedings of ACM-SIGMOD 2005 International Conference on Management of Data (SIGMOD 2005). 2005;658-669
- [10] Brown R. Calendar queues: A fast $O(1)$ priority queue implementation for the simulation event set problem[J]. Communications of the ACM,1988,31(10):1220-1227
- [11] Tang W T, Goh R S M, Thng I L J. Ladder queue: An $O(1)$ priority queue structure for large-scale discrete event simulation[J]. ACM Transactions on Modeling and Computer Simulation, 2005,15(3):175-204
- [12] Onngren R, Riboe R J, Ayani R. Lazy queue: a new approach to implementing the pending event set[J]. International Journal in Computer Simulation,1993,3:303-332
- [13] Sleator D D, Tarjan R E. Self-adjusting binary search trees[J]. Journal of the ACM,1985,32:652-686
- [14] Sleator D D, Tarjan R E. Self-adjusting heaps[J]. SIAM Journal on Computing,1986,15:52-69
- [15] Tarjan R E. Amortized computational complexity[J]. SIAM Journal on Algebraic and Discrete Methods,1985,6:306-318
- [16] Marin M. An Empirical Comparison of Priority Queue Algorithms[R]. Oxford University,1997

2006

- [4] 姜宁康,时成阁.网络存储导论[M].北京:清华大学出版社,2007:28-31
- [5] 严南.计算机应用系统中数据存储系统的可靠性研究[D].昆明:昆明理工大学,2006
- [6] 杨照宏.分布式海量存储系统的可靠性和容错性研究[D].大连:大连海事大学,2007
- [7] 赖莉.数据存储方案分析[J].渝西学院学报:自然科学版,2004,4:43-47
- [8] Lynch M. The Storage Report—Customer Perspectives & Industry Evolution[R]. 2001
- [9] Relex Software Co. Intellect. Reliability: A Practitioner's Guide [M]. 2005;145-151
- [10] 金星,洪廷姬.系统可靠性与可用性分析方法[M].北京:国防工业出版社,2007:12-13