第34卷 第5期 计 算 机 工 程 2008年3月
Vol.34 No.5 Computer Engineering March 2008
·网络与通信· 文章编号：1000—3428(2008)05—0098—03 文献标识码：A 中图分类号：TP393

蓝鲸高可用服务部署系统的研究与设计

汤海鹰[1,2]，冯 硕[1]，杨树庆[1,2]，许 鲁[1]
(1. 中国科学院计算技术研究所，北京 100080；2. 中国科学院研究生院，北京 100039)

摘 要：基于网络存储的特点造成蓝鲸服务部署系统的可用性从根本上依赖于存储服务系统的可用性。该文设计了为客户机提供高可用服务、针对各部件采用不同冗余机制的蓝鲸高可用服务部署系统。对高可用系统的可用性进行了评估。结果表明，与原服务部署系统相比，该系统拥有更高的可用性等级。

关键词：服务部署系统；高可用；失效切换

Research and Design of Bluewhale HA-SonD System

TANG Hai-ying[1,2], FENG Shuo[1], YANG Shu-qing[1,2], XU Lu[1]
(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080;
2. Graduate School, Chinese Academy of Sciences, Beijing 100039)

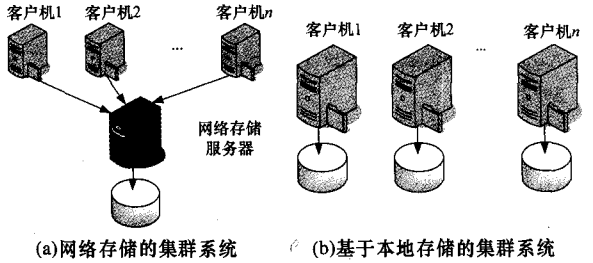
【Abstract】 The availability of SonD system is limited by that of network storage. This paper describes design and implementation of Bluewhale HA-SonD system, which focuses on providing high availability services for client machines and adopts different redundancy mechanisms to each component. The availability evaluation shows that the availability level of HA-SonD system is higher than that of SonD system.

【Key words】 Service-on-Demand(SonD) system; High Availability(HA); failover

蓝鲸服务部署系统(SonD)[1]提供基于网络存储的动态、灵活和快速的服务部署机制。本文针对 SonD 系统采用集中式网络存储造成可用性过于依赖存储服务的问题，为其主要组件设计了不同方案，使系统为用户提供高可用服务。

1 背景介绍

集中式网络存储的系统可用性比传统模式更依赖存储设备。图 1 描述了采用网络存储与本地存储的集群系统的结构区别。假定客户机可用性为 $A_i(i=1,2,\cdots,n)$，网络存储服务器可用性为 A_s，网络存储设备可用性为 A_{sd}，客户机本地存储设备可用性为 A_{cd}。网络存储设备相对本地存储设备未采用额外高可用机制，即 $A_{sd}=\prod_{i=1}^{n}A_{cd}$。单一客户机正常运行时集群系统也能够正常运行。



(a)网络存储的集群系统 (b)基于本地存储的集群系统

图 1 集群系统

基于本地存储与基于网络存储的集群系统可用性为(不考虑网络可用性)：

$$A_{ls}=1-\prod_{i=1}^{n}(1-A_{cd}\times A_i)$$

$$A_{ns}=A_s\times A_{sd}\times(1-\prod_{i=1}^{n}(1-A_i))=A_s\times\prod_{i=1}^{n}A_{cd}\times(1-\prod_{i=1}^{n}(1-A_i))$$

假定 A_i 和 A_s 为 99%，A_{cd} 为 99.99%，表 1 为 2 类系统可用性比较。

表 1 集群系统可用性比较 (%)

客户机个数	A_{ls}	A_{ns}
1	98.990 100	98.000 199
2	99.989 801	98.970 303
3	99.999 897	98.970 204
4	99.999 999	98.960 405

基于本地存储的集群系统可用性等级随客户机增加而提高；基于网络存储的集群系统可用性受存储服务器和存储设备可用性限制，比各部件可用性等级低，大于 2 个客户机的系统可用性随着节点增加反而降低。存储服务可用性是基于网络存储的 SonD 系统提供高可用服务的关键。

2 HA-SonD 系统设计

2.1 总体结构

图 2 是 HA-SonD 系统结构。高可用管理服务器包括提供监控和部署服务的主管理服务器(Master Management Server, MMS)和多个热备从管理服务器(Slave Management Server, SMS)，模式为 Active/Passive。高可用存储服务器中卷管理服务为 Active/Passive 模式、网络存储服务为 Active/Active 模式。主存储服务器(Master Storage Server, MSS)提供卷管理和网络存储服务，多个从存储服务器(Slave Storage Server, SSS)提供网络存储服务。共享存储设备的存储服务器互为备份。服务器的存储设备可靠性通过 RAID 机制保证，不包含在本文范围内。

基金项目：国家“973”计划基金资助项目(2004CB318205)；国家自然科学基金资助项目(60373045)
作者简介：汤海鹰(1977－)，女，博士研究生，主研方向：网络存储，高可用系统；冯 硕，研究实习员；杨树庆，硕士研究生；许 鲁，研究员、博士生导师
收稿日期：2007-03-30 **E-mail**：tanghaiying@nrchpc.ac.cn

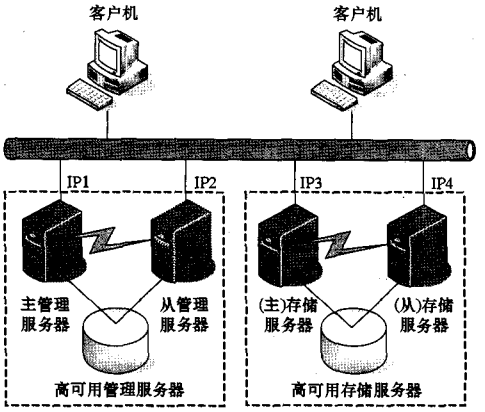


图 2 HA-SonD 系统结构

服务模式不同决定了两种服务器采用不同高可用模式。管理服务器部署服务时间较短，负载较低，而启动多个服务器可能造成数据不一致。存储服务器的卷管理服务类似。存储服务器的网络存储服务运行时间较长，负载较高，可用性也更关键。网络存储服务采用 Active/Active 模式的优点有：(1)多个服务器共同服务可有效使用存储设备；(2)服务器处于 Active 状态可减少失效切换时间；(3)多个服务器为读写不同逻辑卷的不同客户机提供服务不会带来逻辑层写数据冲突。

2.2 模块组成

HA-SonD 系统模块组成如图 3 所示。

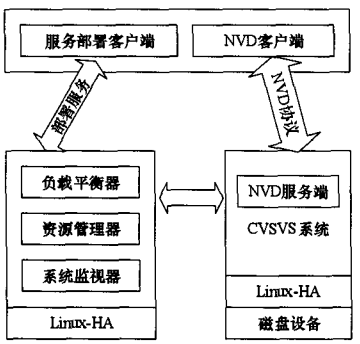


图 3 HA-SonD 系统模块

本系统采用了 3 个关键技术：(1)负载平衡器：在同一逻辑卷可被多个存储服务器导出时分发请求；(2)集群化虚拟共享卷服务 CVSVS 系统[2]：为存储服务失效切换管理底层存储资源；(3)NVD(Network Virtual Disk)数据传输协议[3]：支持数据传输的失效切换，消除服务迁移造成客户机与服务器间连接失效、甚至丢失数据的问题。

2.3 负载平衡器

负载平衡器主要为客户机选择存储服务器提供数据服务。在 SonD 系统使用 Push 模型快照算法基础上，为保证数据完整性，逻辑卷在导出时需遵循两个约束条件：(1)若一个卷被某个服务器导出且可写，则该卷的源及所有快照都应当被该服务器导出；(2)若一个卷被某个服务器导出且可写，则该卷不能被其他服务器导出。最简单的情况中，系统只有一个源卷，其他卷均为其快照卷，那么只需将源卷被所有服务器以只读方式导出、快照卷被相应服务器导出即可。负载平衡器还可采用轮循与负载结合的算法。

2.4 CVSVS 系统

不同存储服务器分别为不同客户机提供服务，它们所导出逻辑卷从用户角度看是不同的。实际上为了实现按需分配、存储虚拟化、备份等，大多数客户机逻辑卷为同一源卷的可写快照。写其中一个卷会移动底层逻辑卷数据和改变内存数据结构，不同服务器上对不同逻辑卷的写可能造成内存数据结构不一致。CVSVS 系统的逻辑卷服务主要功能包括底层物理存储虚拟化、逻辑资源管理、服务节点扩展等。它提供的卷管理和节点扩展功能从底层支持逻辑卷在不同服务器迁移，是存储服务器卷服务失效切换的基础支持。

2.5 NVD 数据传输协议

NVD 是一个适用于高可用服务的数据传输协议。客户端读写操作采用了基于 UDP 协议的 RPC 远程调用机制，传输过程如图 4 所示。

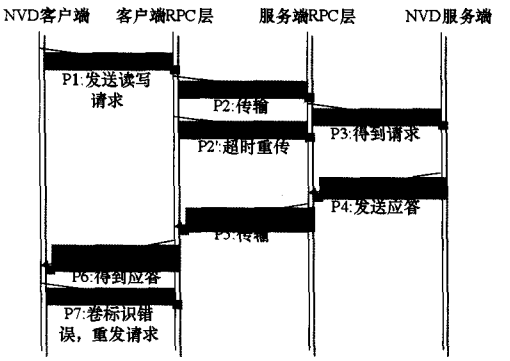


图 4 NVD 数据传输协议

正常读写请求过程为 P1->P2->P3->P4->P5->P6。服务失效切换时，会增加超时重传过程 P2'和卷标识错误重发请求过程 P7，整个处理过程为 P1->P2->P2'->P3->P4->P5->P6->P7->P2…->P6，其间可能经过多次重发。P2'发生在服务器失效和其他服务器替换服务之间，此时发送给服务器的请求表现为超时，底层 RPC 机制保证了超时请求的重传。P7 发生在替换服务器启动 NVD 服务后到导出逻辑卷前接收到客户端请求。由于尚未导出客户端逻辑卷，服务端返回卷标识错误。客户端接收此错误应答后将重发读写请求，直到服务端返回正确应答或重试次数达到上限。

功能和性能测试表明 NVD 能支持对客户端应用透明的服务器失效切换，且性能较高。

3 失效切换和恢复流程

高可用系统核心功能是服务器失效时的服务切换功能与服务器正常工作后的失效恢复功能。HA-SonD 系统包含管理服务器及存储服务器的失效切换和恢复。两类服务器也可能同时失效。

3.1 管理服务器的失效切换和恢复流程

MMS 失效切换流程包括：(1)SMS 监测 MMS 失效；(2)SMS 启用 MMS 的 IP 地址；(3)SMS 启动管理服务。

当原 MMS 恢复时，如果不采用将服务切回原 MMS 的机制，则原 SMS 成为 MMS；否则失效恢复流程包括：(1)SMS 监测到 MMS 恢复；(2)SMS 停止管理服务；(3)SMS 停止使用管理服务 IP 地址；(4)MMS 启动原 IP 地址；(5)MMS 启动管理服务。SMS 失效切换和恢复较简单，不会发生服务迁移，由 MMS 向管理员发警报即可。

3.2 存储服务器的失效切换和恢复流程

存储服务器失效切换流程包括：(1)MSS 或者 SSS 监测对方失效；(2)服务器启用失效服务器 IP 地址；(3)服务器从管理服务器得到失效服务器导出卷信息，通过 NVD 服务导出卷；(4)服务器启动 CVSVS 服务(失效服务器为 SSS 时也需要启动 CVSVS 服务重新读取部分卷内存数据结构)。

存储服务器失效恢复流程包括：(1)替换服务器停止导出

失效服务器的逻辑卷；(2)替换服务器为 MSS 则重启 CVSVS 服务；替换服务器为 SSS 则停止 CVSVS 服务；(3)替换服务器停止使用失效服务器 IP 地址；(4)失效服务器启用 IP 地址；(5)失效服务器导出逻辑卷；(6)失效服务器为 MSS 则启动 CVSVS 服务。

3.3 两类服务器同时失效的流程

主管理服务器与存储服务器可能同时失效。此时管理服务器失效切换与恢复流程与前述相同。存储服务器失效切换时，替换服务器需要从管理服务器得到失效服务器导出卷信息。如管理服务器失效，替换服务器需要等待管理服务器服务正常后才能得到正确信息，增加了存储服务切换时间。存储服务器失效恢复流程没有变化。

4 可用性评估

4.1 SonD 系统可用性

SonD 系统状态包括：

(1)状态 0，管理服务器与存储服务器均正常运行；

(2)状态 1，管理服务器失效，存储服务器正常运行；

(3)状态 2，管理服务器正常运行，存储服务器失效；

(4)状态 3，管理服务器与存储服务器均失效。

假定管理服务器失效率和修复率为 λ_m 和 μ_m，存储服务器失效率和修复率为 λ_s 和 μ_s，2 个服务器都失效时，同一时刻只修复一个服务器，SonD 系统马尔可夫链如图 5 所示。

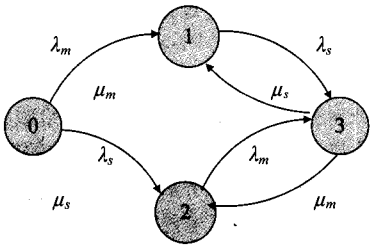


图 5 SonD 系统马尔可夫链

可得 SonD 系统稳态可用性为 $A=\frac{\mu_s}{\lambda_s+\mu_s}$。即 SonD 系统可用性与存储服务器可用性相等。

4.2 HA-SonD 系统可用性

为简化评价过程且保证评价结果的正确性，本文用保守模型和最优模型[4]评价 HA-SonD 系统可用性。保守模型中，高可用管理服务器和高可用存储服务器中任一服务器失效都使系统失效；最优模型中，只有高可用存储服务器失效导致系统失效。系统实际可用性在两者之间。

保守模型中系统可用性 A_c、高可用管理服务器可用性 A_m 和高可用存储服务器可用性 A_s 的关系有 $A_c=A_m\times A_s$。根据文献[8]的双节点可用性可得包含多节点系统可用性 A_m 与 A_s 为

$$A_m=1-[(1-A_{m1})^n+\frac{T_f\times(1-A_{m1})}{3\,600\times MTTR}]=$$

$$1-(\frac{\lambda_m}{\lambda_m+\mu_m})\times[(\frac{\lambda_m}{\lambda_m+\mu_m})^{n-1}+\frac{T_{fm}\times\mu_m}{3\,600}]$$

$$A_m=1-(\frac{\lambda_s}{\lambda_s+\mu_s})\times[(\frac{\lambda_s}{\lambda_s+\mu_s})^{n-1}+\frac{T_{fs}\times\mu_s}{3\,600}]$$

其中，n 为节点个数；A_{m1} 为单个管理服务器可用性；T_{fm} 与 T_{fs} 分别表示高可用管理服务器和高可用存储服务器服务切换时间。存储服务器服务失效切换时间为 3.2 节中的 4 个步骤时间总和。客户机在第 3 个步骤即 NVD 服务导出逻辑卷后就可进入正常数据访问，客户机可视卷服务切换时间为前 3 个步骤时间和加上 NVD 协议重传的附加时间(不考虑 2 类服务器同时失效)。这 4 个时间如表 2 所示。

表 2 客户机监测的服务失效切换时间

步骤	时间/s
心跳监测	~10
IP 地址启用	~2
逻辑卷导出	<10
NVD 协议	~10
总和	~30

心跳监测时间可以配置，一般约为 10 s；IP 地址替换时间较短，为秒级；NVD 服务导出 1 000 个逻辑卷需要 10 s 左右(Intel® Xeon™ 2.4 GHz 虚拟机测试结果)，实际应用中逻辑卷数量少于 1 000 个。NVD 协议重发请求时客户机所需额外时间为 10 s 左右。客户机可视服务停止时间 T_{fs} 为 30 s 左右。假定 $\lambda=\lambda_m=\lambda_s$，$\mu=\mu_m=\mu_s$，$T_{fm}=T_{fs}=30$，高可用管理服务器和高可用存储服务器节点数相同，保守模型系统可用性为

$$A_c=[1-(\frac{\lambda}{\lambda+\mu})\times((\frac{\lambda}{\lambda+\mu})^{n-1}+\frac{\mu}{120})]^2$$

最优模型系统可用性为

$$A_o=1-(\frac{\lambda}{\lambda+\mu})\times((\frac{\lambda}{\lambda+\mu})^{n-1}+\frac{\mu}{120})$$

假定节点平均故障时间分别为 100 h，1 000 h 和 10 000 h，即 λ 取值 0.01，0.001 和 0.000 1，平均修复时间分别为 1 h 和 2 h，即 μ 取值 1 和 1/2，n 为 2，表 3 为 SonD 系统可用性与 HA-SonD 系统保守及最优可用性比较。

表 3 SonD 与 HA-SonD 可用性比较

	λ	A/(%)	Ac/(%)	Ao/(%)
μ=1	0.01	99.009 901	99.963 896	99.981 946
	0.001	99.900 100	99.998 135	99.999 068
	0.000 1	99.990 001	99.999 831	99.999 916
μ=1/2	0.01	98.039 216	99.890 457	99.945 213
	0.001	99.800 399	99.995 877	99.997 938
	0.000 1	99.980 004	99.999 659	99.999 829

数据表明 HA-SonD 系统比 SonD 系统的可用性等级高，达到了系统设计目标。

5 结束语

本文介绍的蓝鲸高可用服务部署系统对蓝鲸服务部署系统可用性瓶颈和单点失效问题进行了分析，为系统关键组成部件设计冗余，改进核心数据传输协议使之支持重传机制，提高了系统可用性。

参考文献

[1] 刘振军, 许 鲁, 尹 洋. 蓝鲸 SonD 服务部署系统[J]. 计算机学报, 2005, 28(7): 1110-1117.

[2] 杨树庆. 集群化的虚拟共享卷服务系统研究[D]. 北京: 中国科学院计算技术研究所, 2006.

[3] Tang Haiying, Feng Shuo, Zhang Huan, et, al. NVD——The Network Virtual Device for HA-SonD[C]//Proceedings of International Workshop on Networking, Architecture, and Storages. Shenyang, China: [s. n.], 2006: 62-169

[4] Reibman A L, Veeraraghavan M. Reliability Modeling: An Overview for System Designers[J]. IEEE Computer, 1991, 24(4): 49-57.

[5] Guo Hui, Zhou Jingli, Li Yue, et al. Design of A Dual-computer Cluster System and Availability Evaluation[C]//Proceedings of the 2004 IEEE International Conference on Networking, Sensing & Control. Taipei, Taiwan, China: [s. n.], 2004: 355-360.