

# 海量下的分布式文件系统测试平台设计与实现

温东新<sup>1</sup>, 王雪娇<sup>1</sup>, 张展<sup>1</sup>, 钱军<sup>1</sup>, 张中兆<sup>2</sup>, 崔忠强<sup>1</sup>

(1. 哈尔滨工业大学 计算机科学与技术学院, 150001 哈尔滨; 2. 哈尔滨工业大学 电子信息与工程学院, 150001 哈尔滨)

**摘要:** 为了对海量存储中的分布式文件系统的容错能力进行测试, 设计并实现了基于海量存储下的自动化分布式文件系统测试平台, 采用从服务层和目录文件层进行故障注入的技术, 其中包括服务器失效和目录文件操作失败等多种故障. 该平台集成了注入故障工具、系统监测工具和工作负载工具, 工作人员可通过平台的主控端进行评测. 通过实验数据表明, 该测试平台能够有效评测不同文件系统的容错性能, 并对不同存储系统的各项技术指标进行比较.

**关键词:** 存储系统; 分布式文件系统; 故障注入; 容错测试

**中图分类号:** TP302

**文献标志码:** A

**文章编号:** 0367-6234(2012)11-0055-04

## Design and realization of test platform for distributed file system under the massive storage

WEN Dong-xin<sup>1</sup>, WANG Xue-jiao<sup>1</sup>, ZHANG Zhan<sup>1</sup>, QIAN Jun<sup>1</sup>, ZHANG Zhong-zhao<sup>2</sup>, CUI Zhong-qiang<sup>1</sup>

(1. School of Computer Science and Technology, Harbin Institute of Technology, 150001 Harbin, China;

2. School of Electronics and Information Engineering, Harbin Institute of Technology, 150001 Harbin, China)

**Abstract:** To evaluate distributed fault-tolerance, an automated test platform is designed and realized based on distributed fault-tolerance for massive storage, which contains a system monitor, a workload injection module and two fault-injection tools based on server level and directory and file operation. This platform can inject different levels faults including server level, directory level and files level, and can finish the test work by the master node. Experimental results show that the platform can compare performance index of the target system with different fault tolerance.

**Key words:** storage system; distributed file system; fault inject; test fault-tolerance

随着现代化信息技术的发展, 存储系统已普遍应用于银行、电信等众多领域. 因此, 对于存储系统的要求也不再仅仅限于容量和性能等方面, 还需从系统容错能力、容灾能力以及可靠性、可用性等方面来优化整个存储系统. 为了保证存储系统的设计质量, 对系统进行可靠性、可用性、容错能力的分析及测试便成为又一研究课题<sup>[1-4]</sup>, 其中故障注入法是常用的测试方法<sup>[7-10]</sup>. 典型的存储系统分为: 应用层、分布式文件系统层、网络驱

动层和磁盘阵列层. 无论故障出现在哪一层, 都会导致系统对外的服务失效, 因此每一层都有相应的容错机制. 在文献[5]中, 针对存储系统的网络层测试平台进行了设计. 本文将研究如何对分布式文件系统的容错能力进行测试, 从而保证存储系统设计的质量<sup>[6]</sup>.

## 1 总体架构设计

本文设计的分布式文件系统容错性能测试平台结构如图1所示. 它包含了工作负载模拟工具、故障注入工具、监测工具以及两个模块: 管理模块和通信模块. 其中管理模块和通信模块将其他各工具模块集成在整个架构下, 测试人员可以在该平台下利用统一的界面接口进行测试.

收稿日期: 2011-11-25.

基金项目: 国家高技术研究发展计划基金资助项目(2009AA01A404); 科技部国际科技合作计划资助项目(2010DFA14400).

作者简介: 温东新(1971—), 女, 副教授、博士;  
张中兆(1951—), 男, 教授, 博士生导师.

通信作者: 温东新, wdongxin@hit.edu.cn.

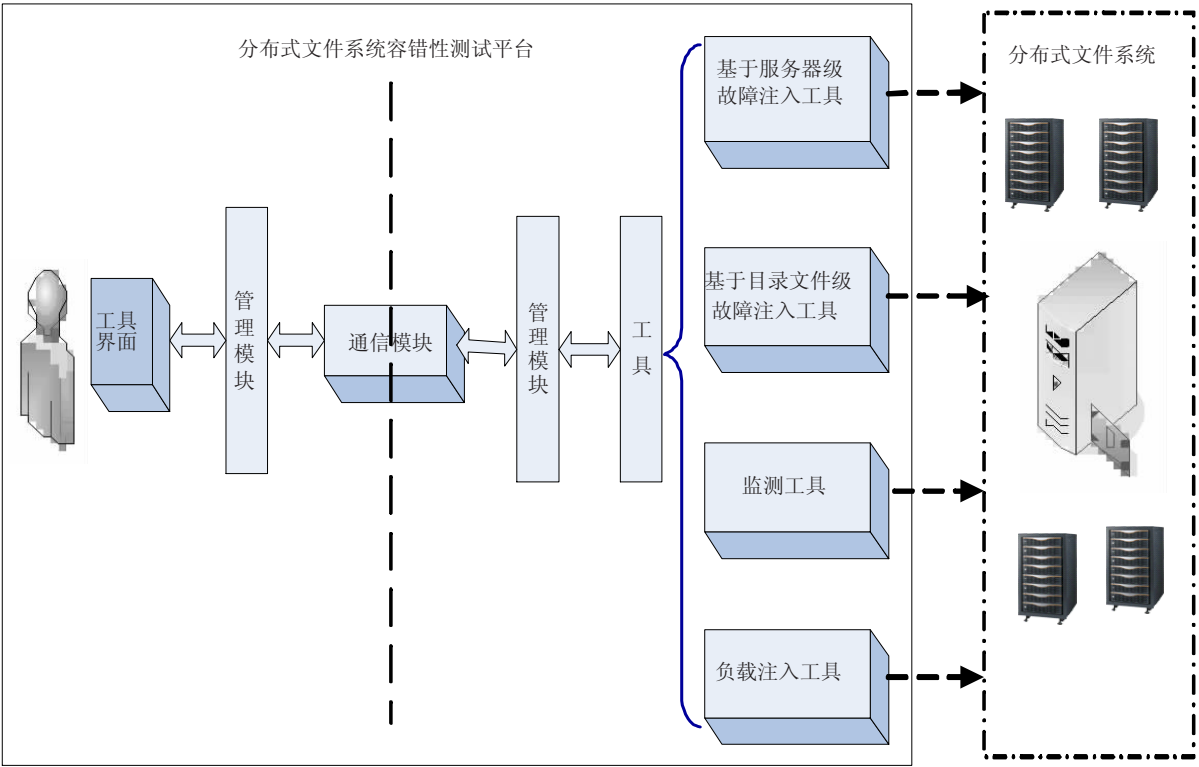


图 1 系统的架构图

由于故障注入对目标系统会产生不可控制的负载,还会影响平台本身的测试工作,故为了测试工作顺利地进行并保证测试得出正确结果,将测试平台设计成 C/S 模式,本文将 C/S 两端分别称为主控端和目标端,在图 1 中用虚线分开. 主控端向用户提供界面接口,而平台中的各种工具的控制和执行则放到目标端,目标端向所有目标系统提供服务,包括负载注入、故障注入以及系统监测.

2 测试平台的设计与实现

本文将使用 Windows 操作系统机器作为主控端,与所需测试的目标系统各个节点分开. 主控端向测试人员提供各工具的界面接口,通过对它的参数进行配置,发出相应的指令,例如为工作负载模拟工具配置用例参数可发出模拟工作指令,配置故障注入工具用例参数可发出故障注入指令,将监控指令发给监控工具.

目标端,即为目标系统的各节点,主要执行故障注入、系统监测和工作负载模拟. 管理模块和通信模块安装在目标系统的客户端上,该客户端需要保持无故障状态.

1)故障注入. 平台的两个注入工具分别基于服务层和目录文件层. 服务节点中的故障注入工具通过向元数据服务器注入故障,使服务器工作在故障环境下,通过平台的监测工具系统监测和

工作负载模拟,测试目标系统针对该故障的容错能力.

客户端的目录文件层注入工具通过注入文件打开失败、读写失败、目录打开/创建/删除失败等故障,评测出目标系统针对此类故障的容错性能.

2)系统监测. 节点中的系统监测包括服务器端和客户端的监测,监测故障注入后的运行结果,对目标文件系统的状态查询及系统的磁盘分区查询等,通过监测工具来显示监测结果.

3)工作负载模拟. 工作负载调度模块将 3 种典型的工作负载 FDTREE、IOZONE 和 FDOS 进行管理. FDTREE 是开源测试元数据的吞吐率的测试工具;IOZONE 是读写性能的测试工具;FDOS 是本文开发的目录文件操作的集合脚本. 工作负载模拟为测试分布式目标文件系统提供模拟的工作负载,在工作负载执行过程中,分析目标系统在故障影响下负载的执行情况,从而反映出目标系统对该故障的容错能力.

2.1 服务器层故障注入工具

服务器层故障注入工具结构如图 2 所示. 服务器级故障注入工具分布于每个服务器节点. 测试人员在主控端的测试界面下配置故障用例,发出注入故障指令. 目标端根据主控端的指令执行相应的故障注入. 图 2 中的管理模块和通信模块是测试平台的公共模块.

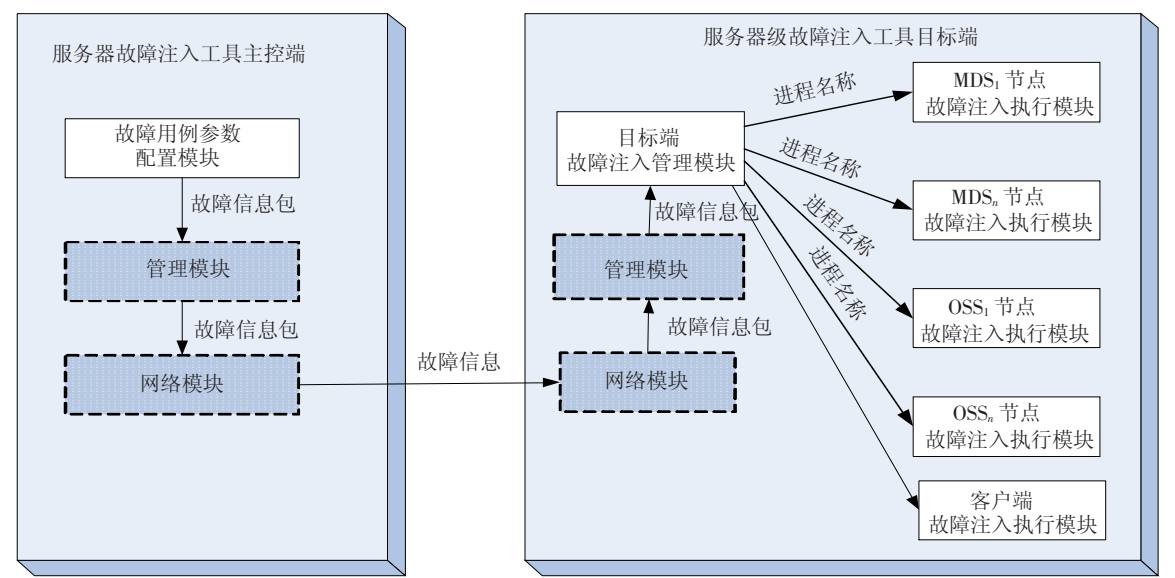


图 2 服务器级故障注入工具架构图

- 1)故障用例参数配置模块. 此模块主要提供参数配置界面.故障参数包括:故障位置、节点 IP 地址和故障进程名称.配置故障位置是确定故障注入的对象,即服务器的位置.配置节点 IP 地址是确定故障节点的 IP.该工具利用杀掉目标系统服务进程的方式模拟故障注入,所以要给出涉及到的故障进程名称.
- 2)目标端故障注入调度模块.该模块在分布式文件系统目标端,作用是管理各个节点故障注入执行.当收到来自管理模块的故障信息后,根据故障位置来调动相应节点的注入执行模块,根据故障节点 IP 地址访问相应节点.
- 3)服务器故障注入执行模块.在系统的各个服务器节点中,都安装有实现故障注入执行的模块.该模块处于工具调度的最底层,完成最后故障注入的操作.

执行过程如图 3 所示,先根据客户端的故障注入管理模块查询进程的相应信息,再从得到的信息中提取出进程号,然后根据进程号,调用命令“kill-9 进程号”,完成杀进程操作,目的是使所在节点的服务进程失效.

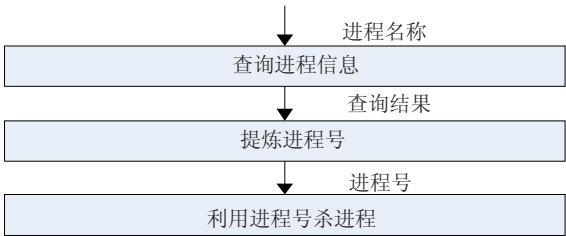


图 3 服务器故障注入的执行过程

2.2 目录文件层故障注入工具

目录文件层的故障注入工具结构如图 4 所示.该工具利用动态可加载的模块内核技术,开发

一个可加载内核模块,在平台启动时,可将该模块加载到内核中.当故障执行时把相关信息传递给该模块.该工具可分为用户空间、内核空间和 /proc 缓存 3 部分.

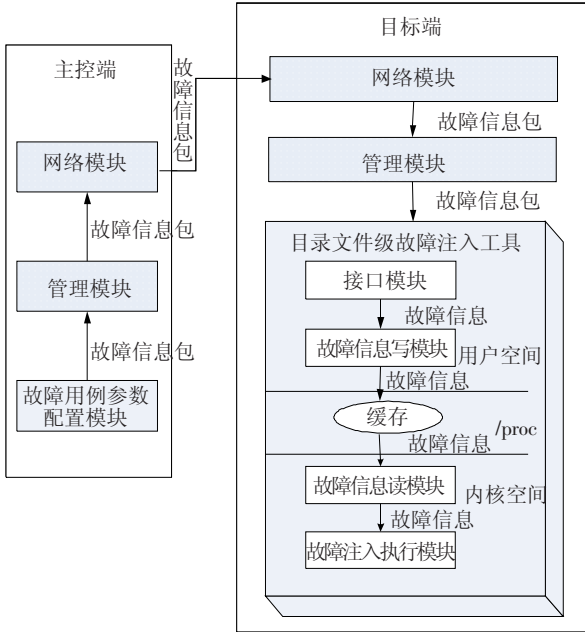


图 4 目录文件级故障注入工具架构图

- 1)用户空间. 用户空间包含两个模块:故障信息写模块和接口模块.当接口模块收到平台管理模块传来的故障信息包后除掉信息包的头部信息,并将包信息传给故障信息写模块,这个包信息就是故障信息,它包括故障位置及故障名称.此后故障信息写模块把收到的故障信息写到 /proc 缓存中.
- 2)/proc 缓存.该缓存部分主要是利用 Linux 虚拟文件系统 /proc 来完成./proc 文件系统位于内存,不占硬盘.本课题为此建立了故障注入的位置缓存池和名称缓存池,将利用故障信息写模块

向两个缓存池中写入相应的信息,在内核空间的故障信息读模块能够自动读取该信息.

3)内核空间. 内核空间包含故障信息读模块和注入执行模块. 其中故障信息读模块通过开启线程,从缓存池中不断地读取信息,然后发给故障的注入执行模块,实现最终故障注入.

### 3 结 果

由于容错性能与环境配置关系密切,本文将项目研究中的两种典型的文件系统 Adfs 和 Bdfs (根据研究工作要求,这里只用符号表示两种不同的文件系统)测试环境的软硬件配置及各节点的软硬件配置列表如表 1 和表 2 所示.

本文测试了 Adfs 和 Bdfs 两种文件系统在服务器层和目录文件层的故障容错能力. 实验中故障集包括:MDS 失效、OSS 失效、Dir\_Open\_Erro、Dir\_Create\_Erro、Dir\_Del\_Open、File\_Open\_Erro、File\_Read\_Open 和 File\_Write\_Erro 等. 实验表明两种不同文件系统对上述故障具有不同的容错能力. 依据性能降级的情况以及故障所导致对外服务的延迟时间,能够得出对于服务器层的故

障,Bdfs 的容错能力优于 Adfs,对于目录文件层的故障,二者的容错能力是相近的. 在这里通过表 3 给出了 MDS 失效故障导致的测试结果.

| 表 1 Adfs 文件系统测试环境硬件配置 |          |       |                                            |
|-----------------------|----------|-------|--------------------------------------------|
| 配置                    | 内核       | 内存/GB | CPU                                        |
| client 端              | 2. 6. 18 | 32    | Intel(R) Xeon(R) CPU X5660 @ 2. 80 GHz 2 颗 |
|                       | 2. 6. 18 | 32    | Intel(R) Xeon(R) CPU X5660 @ 2. 80 GHz 2 颗 |
| MDS 端                 | 2. 6. 18 | 32    | Intel(R) Xeon(R) CPU X5660 @ 2. 80 GHz 2 颗 |
|                       | 2. 6. 18 | 32    | Intel(R) Xeon(R) CPU X5660 @ 2. 80 GHz 2 颗 |
| OSS 端                 | 2. 6. 18 | 32    | Intel(R) Xeon(R) CPU X5660 @ 2. 80 GHz 2 颗 |
|                       | 2. 6. 18 | 32    | Intel(R) Xeon(R) CPU X5660 @ 2. 80 GHz 2 颗 |

| 表 2 Bdfs 文件系统测试环境硬件配置 |          |       |                                           |
|-----------------------|----------|-------|-------------------------------------------|
| 配置                    | 内核       | 内存/GB | CPU                                       |
| 客 户 端                 | 2. 6. 18 | 12    | Intel(R) Xeon(R) CPU E5620@2. 40 GHz 16 核 |
| MDS 端                 | 2. 6. 32 | 12    | Intel(R) Xeon(R) CPU E5620@2. 40 GHz 16 核 |

表 3 Adfs 和 Bdfs 在 MDS 失效后服务延迟时间和每秒删除文件数比较表

| 测试系统 | 延迟时间/s |       |       |       |       |    | 删除文件数量 |       |       |       |       |     |
|------|--------|-------|-------|-------|-------|----|--------|-------|-------|-------|-------|-----|
|      | 第 1 次  | 第 2 次 | 第 3 次 | 第 4 次 | 第 5 次 | 平均 | 第 1 次  | 第 2 次 | 第 3 次 | 第 4 次 | 第 5 次 | 平均  |
| Adfs | 70     | 69    | 67    | 66    | 65    | 68 | 455    | 450   | 456   | 466   | 456   | 456 |
| Bdfs | 7      | 8     | 8     | 7     | 7     | 7  | 688    | 690   | 687   | 683   | 687   | 687 |

### 4 结 论

本文设计并实现了基于海量存储下的自动化分布式文件系统测试平台. 平台集成了注入故障工具、系统监测工具和工作负载工具,主要采用故障注入技术对海量存储中的不同分布式文件系统的容错能力进行测试. 注入的故障包括服务器失效和目录文件操作失败等多种故障. 利用测试平台能够对不同的文件系统的容错能力进行测试比较,通过分析测试结果,更有利于完善文件系统的设计.

### 参考文献:

[1] HSUEH Mei-Chen, TSAI T K., IYER R K. Fault injections techniques and tools [J]. IEEE Computer, 1997, 30(4): 75-82.

[2] FENG Dan. The next generation of Internet information storage and core technology research organization model [C]//China Computer Federation Proceedings CCFP 0008. [S.l.]:[s.n.], 2007.

[3] 张英, 张益, 王冀鲁. 基于框图法的网络存储系统可靠性分析[J]. 计算机科学, 2010, 37(6): 102-106.

[4] 杨宇航, 冯允成. 基于仿真的复杂系统可靠性、可用性和 MTBF 评估文献综述[J]. 系统工程理论与实践, 2003, 23(2): 80-85.

[5] 陈宇龙, 温东新, 张展, 等. 基于 SCSI 的故障注入工具设计与实现[J]. 计算机工程, 2011, 37(22): 222-225.

[6] 王雪娇. 分布式文件系统容错能力测试平台的设计与实现[D]. 哈尔滨: 哈尔滨工业大学, 2011.

[7] 王雪娇. 基于 Linux 虚拟文件系统故障注入器的设计与实现[J]. 计算机研究与发展, 2010, 47(S1): 122-126.

[8] 潘清, 庄泽南, 王霄军, 等. 测量海量存储系统可用性的故障注入法[J]. 计算机工程与应用, 2010, 46(36): 85-94.

[9] 仇俊峰, 洪炳熔, 乔永强. 基于软件方法故障注入系统[J]. 哈尔滨工业大学学报, 2006, 38(6): 873-876.

[10] 朱俊杰, 梁阿磊, 马汝辉. 海量存储系统高可用性研究[J]. 计算机应用与软件, 2012, 29(7): 81-82, 86.

(编辑 张 红)