GPFS 文件系统的安装配置与维护

张新诺,王　彬
(国家气象信息中心,北京 100081)

摘　要:GPFS 是一款成熟的并行文件系统,该系统具有稳定性好、扩展性高、数据读写速度快等特点,广泛应用于 AIX 和 Linux 操作系统的服务器集群中。虽然 GPFS 文件系统稳定性高,但 GPFS 文件系统是依托在操作系统和存储介质中的软件,当操作系统或存储介质发生故障时,也会造成 GPFS 文件系统无法正常使用。在处理 GPFS 文件系统故障时,需要使用相关命令并配合不同的参数对文件系统的相关内容进行配置和创建。文中的主要目的是为了说明 GPFS 文件系统相关命令及命令中各项参数的作用和使用方式,能够使维护人员更好地了解和维护 GPFS 文件系统。同时,也对文件系统在 Linux 平台中出现问题时如何有效使用 GPFS 命令解决故障,如何有效保护数据安全和数据完整性,及如何恢复文件系统的正常运行等方面进行探讨。

关键词:并行文件系统;GPFS;NSD;重建文件系统
中图分类号:TP302.1　　**文献标识码**:A　　**文章编号**:1673-629X(2018)05-0174-05
doi:10.3969/j.issn.1673-629X.2018.05.039

Installation Configuration and Maintenance of GPFS

ZHANG Xin-nuo, WANG Bin
(National Meteorological Information Center, Beijing 100081, China)

Abstract: GPFS (general parallel file system) is a mature parallel file system with the characteristics of high stability and expansibility, fast data reading and writing and many more, which is widely used in the server cluster of AIX and Linux operating system. Although the GPFS has high stability, it is a software which is run in the operating system and storage media, and when the operating system or storage media fails, it is not used properly. Dealing with the failures of GPFS, we need to configure and create related content of the file system with the relevant commands and different parameters. The main purpose in this paper is to explain the function and usage of the parameters in the GPFS, to make the maintenance staff understand and use the GPFS better. At the same time, we also discuss how to use GPFS command to solve the problem, how to effectively protect data security and data integrity, and how to restore the normal operation of file system, when the file system is in the Linux platform.

Key words: parallel file system; GPFS; NSD; rebuild file system

0　引　言

随着信息化进程的推动,各行各业所需的相关信息量越来越多,需要更高效更安全的数据存储环境。在并行存储迅速发展的情况下,由于性能优势而备受国内各企事业单位信赖的 GPFS 系统得到了广泛应用。

IBM 公司的 GPFS 文件系统全称为 general parallel file system(通用并行文件系统),是 IBM 公司开发并生产的一种并行文件系统,普遍应用于服务器集群系统中[1]。GPFS 文件系统为集群中的节点提供统一的数据存储空间,并允许集群中任何一个节点同时访问相同的数据。简单来说,GPFS 是一个高性能、可共享磁盘的并行文件系统[2]。

1　GPFS 文件系统的特性

GPFS 文件系统为服务器集群提供高性能的数据访问,该文件系统允许数据被集群中多个节点同时、高效的访问。大多数现有的文件系统是专为单一服务器环境提供服务的,添加更多文件服务器并不会提高文件系统的性能。GPFS 文件系统将独立的数据分块,

收稿日期:2017-05-31　　**修回日期**:2017-09-13　　**网络出版时间**:2018-02-08
基金项目:科技部公益性行业专项(气象)科研专项(GYHY201106022,GYHY201306062)
作者简介:张新诺(1984-),男,工程师,硕士,研究方向为并行计算;王　彬,正研级高级工程师,博士,研究方向为高性能计算应用及技术开发、气象信息化设计等。

并存放在多块硬盘中,以并行的方式进行数据的输入和输出,能够为服务器提供高性能的数据服务[3-4]。GPFS 提供的其他功能包括高可用性、支持异构集群、灾难恢复、安全性、数据管理接口(DMAPI)、分级存储管理(HSM)和信息生命周期管理(ILM)[5]。

2　GPFS 文件系统的配置

GPFS 文件系统的安装较为简便,对应不同的操作系统安装其相应的安装包即可。安装完成后,不同的操作系统其对应的配置方式略有不同。文中以 SUSE 10 为例,简述文件系统的配置方式。

2.1　配置节点文件

配置节点文件主要是确定该 GPFS 文件系统可用于集群的范围,并确定文件系统的管理节点和仲裁节点的位置,配置如下:

root@ hs21-1 [/root]
vim /usr/lpp/mmfs/nodef
hs21-1. site:quorum
hs21-2. site
hs21-3. site
x3650-01. site:quorum-manager
x3650-02. site:quorum-manager

在 GPFS 文件系统中,至少需要一个仲裁节点(quorum),用于集群间的通信及数据完整性检查。由于该集群中节点较多,故设置三个仲裁节点,当任何一个仲裁节点出现问题时,集群节点仍能和其他的仲裁节点保持通信,保证 GPFS 文件系统仍能正常运行[6-7]。若多个仲裁节点发生故障,则集群节点无法正常通信,此时 GPFS 文件系统将不可用。在 hs 命名为刀片服务器,x3650 服务器为普通 X86 机架服务器,为了保证文件系统的安全性,将两台 x3650 机架服务器定义为仲裁节点(quorum)和管理节点(manager),即 quorum-manager 类型节点。该种类型节点用于管理集群的配置及文件系统监控等方面[8]。

2.2　建立 GPFS 集群

将 hs21-1. site 作为 GPFS 集群的主管理者,在该节点上进行 GPFS 集群建立的操作。操作命令为:

mmcrcluster -C 集群名 -U 域名 -N 各节点名 -p 主 NSD 服务器 -s 备 NSD 服务器

具体命令如下:

root@ hs21-1 [/root]
#mmcrcluster -C hs21. cma. GPFS -U hs21. cma. GPFS -N /usr/lpp/mmfs/nodef -p hs21-1. site -s x3650-02. site

该命令中各参数含义如下:

-C hs21. cma. GPFS:设定集群名称为 hs21. cma. GPFS
-U hs21. cma. GPFS:设定域名为 hs21. cma. GPFS
-N /tmp/mmfs/nodef:指定各节点的文件名
-p hs21-1. site 指定主 NSD:服务器为 hs21-1. site
-s x3650-02. site 指定备 NSD:服务器为 x3650-02. site

命令执行完成后,执行 mmlscluster 命令检查集群建立情况,结果见图 1。

```
root@hs21-1 [ /root ]
# mmlscluster

GPFS cluster information
========================
  GPFS cluster name:         hs21.cma.gpfs
  GPFS cluster id:           726257272784766060
  GPFS UID domain:           hs21.cma.gpfs
  Remote shell command:      /usr/bin/ssh
  Remote file copy command:  /usr/bin/scp

GPFS cluster configuration servers:
-----------------------------------
  Primary server:    hs21-1.site
  Secondary server:  x3650-02.site

 Node  Daemon node name    IP address    Admin node name    Designation
=================================================================================
   1   hs21-1.site         10.*.*.*    hs21-1.site          quorum
   2   hs21-2.site         10.*.*.*    hs21-2.site
   3   hs21-3.site         10.*.*.*    hs21-3.site
   4   x3650-01.site       10.*.*.*    x3650-01.site        quorum-manager
   5   x3650-02.site       10.*.*.*    x3650-02.site        quorum-manager
```

图 1　GPFS 文件系统建立情况

2.3　配置并创建 GPFS 共享磁盘

2.3.1　建立 NSD(network shared disk)配置文件

该文件系统是创建在由 36 块硬盘组成的存储介质中,其中每 3 块硬盘组建 RAID 5 磁盘阵列[9-10]。可用 fdisk -l 查看各硬盘信息,显示结果如下:

root@ hs21-1 [/root]
#fdisk - l
Disk /dev/sdal: 1998. 9 GB,1998985153536 bytes
255 heads, 63 sectors/track,243029 cylinders
Units = cylinders of 16065 * 512 = 8225280 bytes
……

根据硬盘信息建立 NSD 配置文件,配置文件的文件名可根据个人习惯命名,文中将其命名为 DescFile,用于 NSD 的划分。NSD 配置文件内容格式为:磁盘名:主节点名:备节点名:磁盘类型:失效组别:NSD 名:存储池:(“:”为必须内容)。

根据命令格式编辑 DescFile 文件:

root@ hs21-1 [/root]
viDescFile
/dev/sdal:hs21 - 1. site:x3650 - 02. site:dataAndMetadata:4001:ft01_nsd1::
/dev/sdam:hs21 - 1. site:x3650 - 02. site:dataAndMetadata:4001:ft01_nsd2::
……(省略其他 NSD 硬盘)
/dev/sdbu:hs21 - 1. site:x3650 - 02. site:dataAndMetadata:4001:ft01_nsd36::

该文件中各字段具体解释如下:

(1)/dev/sdbu:代表硬盘名称,通过 fdisk -l 命令获得,不同系统对应的硬盘名称略有不同。

(2) hs21-1. site:代表 NSD 的主 I/O 节点,该节点名称根据文件系统的实际情况配置。

(3) X3650-02. site:代表 NSD 的备 I/O 节点。无该节点可以不填。

(4) dataAndMetadata:代表磁盘类型。NSD 磁盘根据数据类型可以分为四种,分别为 dataAndMetadata、dataOnly、metadataOnly 和 descOnly。

GPFS 需要保存两种类型的数据,即 data 和 metadata。metadata(元数据)是用于 GPFS 自身索引数据以及内部配置信息。这部分元数据只能保存在 dataAndMetadata 或者 metadataOnly 类型的磁盘中。dataAndMetadata 说明该磁盘既可以存放元数据,也可以存放其他数据。在一些对元数据访问要求非常高的系统中,推荐使用 Flash 单独存放 GPFS 的元数据。在这种情况下,Flash 里的磁盘就设置为 metadataOnly,其他的磁盘就设置为 dataOnly。descOnly 类型的磁盘仅用于存放文件系统副本,并且在发生故障时可根据第三失效组恢复相关配置。一般来说,dataAndMetadata 为 NSD 磁盘的默认类型[11-13]。

(5)4001:代表失效组(FailureGroup),可以不填,默认为 4001。失效组主要用于定义一组来自于同一存储系统或者有一定隔离效果(如同一存储中的同一个 RAID)的磁盘[14]。如果启用 GPFS 的 Replica(复制)功能,GPFS 会把同一个数据块的 2 个或者 3 个 replica 放置在不同的 FailureGroup 里。这样的话,同一个 FailureGroup 里不管坏多少个磁盘,都不会影响数据访问。FailureGroup 的数值本身没有具体含义,主要为了区分不同失效组,数值相同的磁盘属于同一个失效组。如果启用 GPFS 的 replica 功能,每个数据块会多占用一倍(replica = 2)甚至两倍(replica = 3)的磁盘空间。一般而言,如果存储系统在硬件上已经保证了冗余,比如 RAID 以及多路径到 SAN 交换机,磁盘失效的概率已经很低,一般情况没必要启用 Replica 功能。

(6) ft01_nsd * :代表 NSD 盘的名称,可根据需要自行命名。

(7)命令最后一位代表存储池,如不填代表系统默认的存储池。

2.3.2 创建 GPFS 所需的 NSD 盘

编辑 DescFile 完成后,执行 mmcrnsd -F 命令,即可生成 NSD 盘。命令格式为:mmcrnsd -F NSD 配置文件。

命令执行如下:

```
root@ hs21-1 [/root]
#mmcrnsd -F DescFile
mmcrnsd:Processing disk sdal
……
mmcrnsd:Propagating the cluster configuration data to all
affected nodes. This is an asynchronous process.
```

命令执行完成后,可执行命令 mmlsnsd -L 检查 NSD 盘的创建情况。

2.4 创建文件系统

NSD 盘创建完成后,需执行 mmstartup -a 启动 GPFS。只有启动 GPFS 后,才能继续进行文件系统的创建。创建文件系统的命令为:

mmcrfs 文件系统设备名“NSD 盘名” -T 文件系统挂载点 - A yes/no -B 数据块大小

具体命令如下:mmcrfs /dev/fs1 " ft01_nsd1;ft01_nsd2;…(中间略)…;ft01_nsd36" -T /GPFS/fs1

命令中各字段具体解释如下:

/dev/fs1 为文件系统设备名,创建文件系统时,系统会在集群所有节点自动创建。在 4.2.1 版本及更新的版本中,GFPS 在 Linux 中将不会在/dev 目录下生成文件系统设备名,因此在 Linux 版本的 mount 命令中也不会出现/dev 的前缀。

“ft01_nsd1;……ft01_nsd36”为前文创建的 NSD 盘。也可使用创建 NSD 盘的 DescFile 文件,命令为:mmcrfs /dev/fs1 - F /root/DescFile -T /GPFS/fs1。

-A 表示开机是否自动加载挂载点,默认是 no,命令中可以不使用。

-B 表示数据块大小,默认是 256K,如果不加-B 则表示使用默认数据块大小,文件系统创建后不可更改。数据块的大小选择与应用程序下发的 I/O 的大小接近时,GPFS 的性能较好。GPFS 本身有很多机制来适应不同的 I/O 大小。当应用程序下发的 I/O 大小不是很清楚,或者很复杂时,可以选择 1 MB 来折中。

2.5 挂载文件系统

执行挂载文件系统命令挂载文件系统,命令为:mmmount 文件系统名。文件系统挂载完成后,可执行 df 命令查看文件系统情况。如能够显示所建文件系统的路径和名称,说明文件系统已经建设完成。具体执行命令为:mmmount fs1。

3 GPFS 的管理与维护

3.1 GPFS 启动与关闭

(1) mmstartup:为启动文件系统的命令,根据添加不同参数可以启动单节点的文件系统,也可启动所有集群文件系统。

单节点文件系统启动方式:执行 mmstartup 命令。

集群文件系统启动方式:执行 mmstartup -a 命令。

(2) mmshutdown:为关闭文件系统命令,同样根据添加不同参数可以关闭单节点文件系统,也可关闭所

有集群文件系统。

单节点文件系统关闭方式:执行 mmshutdown。

集群文件系统关闭方式:执行 mmshutdown -a,其执行结果显示的时间会比关闭单节点时间略长。

通常在重启节点时,要先关闭 GPFS 文件系统等常用软件,避免节点重启完成后 GPFS 软件出现故障。

3.2 GPFS 文件系统故障检查处理

通过长时间使用和维护 GPFS 文件系统,发现 GPFS 文件系统出现的故障多为硬件故障,因此文中主要从硬件角度处理 GPFS 故障。当 GPFS 文件系统出现建故障时,需要通过查看集群 GPFS 状态、磁盘状态、文件系统故障范围等多方面来确定故障原因。

3.2.1 查看集群节点状态

通过命令 mmgetstate - a 可查看集群各节点 GPFS 状态,执行结果如下:

```
root@ hs21-1 [/root]
#mmgetstate -a
Node number   Node name          GPFS state
------------------------------------------
      1       hs21-1             active
      3       hs21-3             active
mmgetstate: The following nodes could not be reached:
hs21-2. site
```

从执行结果可以看到,节点 hs21-2. site 出现故障无法加入到集群中,其他节点 GPFS 都是健康的 active 状态,若 GPFS state 显示其他状态,如 GPFS stat 显示 down 时,登录到相关节点查看 GPFS 是否被关闭。检查/var/adm/ras 目录下的 mmfs. log. latest 文件内容,查看是否有 GPFS 关闭的信息,如执行如下命令:

```
root@ mdss-zc2 [ /var/adm/ras ]
# tail -500mmfs. log. latest
……
Wed May 17 00:40:42 GMT 2017: mmremote: Completing GPFS shutdown ...
```

如结果所示,可知该节点 GPFS 被执行 shutdown 操作,可用 mmstartup 命令尝试能否重启 GPFS 文件系统。若 mmfs. log. latest 文件中没有 GPFS 文件系统的 shutdown 信息,则 GPFS 文件系统进程被关闭或由于其他原因造成文件系统无法正常工作,可先执行 mmshutdown,再执行 mmstartup 尝试重启文件系统。如果无法正常启动,可用其他命令检查物理硬盘等其他故障原因。

3.2.2 查看 NSD 状态检查物理硬盘

在使用 GPFS 文件系统过程中,有时会发生文件系统出现挂起状态,从而导致文件系统不可用,如在执行 df 命令时,显示“df:/dev/fs1:Stale NFS file handle”,说明 fs1 文件系统被挂起。此时可检查文件系统对应 NSD 是否出现故障。

查询文件系统 NSD 命令:mmlsdisk 文件系统名。

执行命令见图 2。

```
root@hs21-1 [ /root ]
# mmlsdisk fs1
disk         driver   sector failure holds    holds                            storage
name         type       size   group metadata data   status        availability  pool
------------ -------- ------ ------- -------- ----- ------------- ------------ ------------
……
ft01_nsd4    nsd         512    4001 yes      yes   ready         down         system
ft01_nsd5    nsd         512    4001 yes      yes   ready         down         system
ft01_nsd6    nsd         512    4001 yes      yes   ready         down         system
```

图 2　查看文件系统 NSD 状态

通过查看可以看出,该文件系统中有三个 NSD 出现故障,显示出 down 的状态。当 NSD 出现 down 的状态时,说明系统中与 down 状态 NSD 相对应的物理硬盘出现故障。根据硬盘状态检查该硬盘所在 RAID 状态。如果 RAID 完好,只需更换故障硬盘,重新挂载文件系统即可;如果 RAID 出现故障,则可能需要重建文件系统。

当存储系统中的 RAID 出现故障,首先要查看 3 块物理硬盘出现故障的原因,尝试能够将故障 NSD 的状态改为 up。如果 NSD 状态能够恢复,则说明不是所有的硬盘都存在物理故障,文件系统中的数据也不会出现丢失的现象,此时,文件系统也可以被恢复并被挂载。恢复 NSD 状态可使用如下命令:

mmchdisk 文件系统名 start -d “故障 NSD1;故障 NSD2;……;故障 NSDn”

具体操作命令为:mmchdisk fs1 start -d “ft01_nsd4;ft01_nsd5;ft01_nsd6”。

执行完成后,检查 NSD 状态。如果 NSD 状态都是 UP 状态,则可执行 mmumount fs1 和 mmmount fs1 来恢复文件系统的正常运行。

若 NSD 状态仍是 down 或者 unreacoverd,则表示文件系统仍存在问题,文件系统中的数据可能已经无法恢复。

在故障状态下,故障 NSD 已经无法与正常 NSD 同步数据,此时文件系统可以挂载,但无法对文件系统中的内容进行操作。为了恢复文件系统的正常运行,

需要重建文件系统。重建文件系统会造成文件系统中的数据丢失,因此需要尽量备份文件系统数据。

为了尽可能备份文件系统中的数据,可以先屏蔽故障 NSD,再尝试以只读模式挂载文件系统。由于文件系统缺少 3 个 NSD,文件系统中的数据已经不完整,在备份数据过程中会出现数据不完整情况[15-19]。屏蔽故障 NSD 和以只读模式挂载文件系统命令如下:

屏蔽故障 NSD:mmfsctl 文件系统名 exclude -d "故障 NSD1;故障 NSD2;故障 NSD3"。

以只读模式挂载文件系统:mmmount 文件系统名-o rs。

3.2.3 重建 GPFS 文件系统

数据备份完成后,需要重建文件系统才能使文件系统恢复正常工作。重建文件系统时,文件系统中所有数据都将丢失。重建文件系统的步骤:删除故障文件系统,删除故障文件系统 NSD,重建 NSD,重建文件系统。具体命令如下:

(1)删除故障文件系统。

mmdelfs -p 故障文件系统名:删除故障文件系统,为重建系统做准备。

(2)删除故障文件系统 NSD。

mmdelnsd "NSD1;NSD2;……NSDn":更换故障硬盘后,为了重建故障文件系统,需要将该文件系统所有 NSD 硬盘初始化,因此需要先删除现有 NSD,再进行 NSD 和文件系统的重建。

(3)重建 NSD 和重建文件系统。

具体操作可参考前文创建 NSD 和创建文件系统时的步骤,完成相关的操作后,即可重新使用该文件系统。

4 结束语

目前,GPFS 并行文件系统广泛应用于各企事业单位。随着信息化的发展,各企事业单位需要结合自身情况来配置使用 GPFS 并行文件系统,保证数据能够被高效、安全地使用。因此,为了更加高效、稳定地使用各种数据,应该继续深入探讨 GPFS 并行文件系统相关内容,为更好地使用 GPFS 文件系统打下基础。

参考文献:

[1] 王 鸥,赵永彬. GPFS 共享文件系统在企业门户系统中应用的研究[J]. 电脑知识与技术,2015,11(10):15-17.

[2] 张志坚,伍光胜,孙伟忠,等. IBMFlexP460 高性能计算机系统及气象应用[J]. 现代计算机,2016(9):51-55.

[3] 庞丽萍,何飞跃,岳建辉,等. 并行文件系统集中式元数据管理高可用系统设计[J]. 计算机工程与科学,2004,26(11):87-88.

[4] 肖 伟,赵以爽. 并行文件系统简介及主流产品对比[J]. 邮电设计技术,2012(7):31-36.

[5] 杨 昕. GPFS 文件系统原理和模式 IO 优化方法[J]. 气象科技,2006,34:27-30.

[6] 张 玺. 并行文件系统下数据迁移功能的实现[J]. 北京信息科技大学学报:自然科学版,2012,27(5):77-80.

[7] 解宝琦,王金国. 构建 CentOS+GPFS 集群[J]. 网络安全和信息化,2017(2):85-88.

[8] 叶雅泉. GPFS 在省级通信系统中的应用[J]. 移动信息,2016(6):113.

[9] SCHMUCK F,HASKIN R L. GPFS:a shared-disk file system for large computing clusters[C]//Proceedings of the conference on file and storage technologies. Berkeley,CA,USA:USENIX Association,2002:231-244.

[10] PLANK J S. The Raid-6 Liber8Tion code[J]. International Journal of High Performance Computing Applications,2009,23(3):242-251.

[11] CAULFIELD A M,SWANSON S. QuickSAN:a storage area network for fast,distributed,solid state disks[J]. ACM SIGARCH Computer Architecture News,2013,41(3):464-474.

[12] 沈 瑜,孙 婧,李 娟. 中国气象局高性能计算机系统高可靠性设计[J]. 信息安全与技术,2013,4(6):42-45.

[13] JONES T,KONIGES A E,YATES R K. Performance of the IBM general parallel file system[C]//Proceedings of the 14th international symposium on parallel and distributed processing. Cancún,Mexico:IEEE,2000:673-681.

[14] LIU Gia-Shie. Three m-failure group maintenance models for M/M/N unreliable queuing service systems[J]. Computers & Industrial Engineering,2012,62(4):1011-1024.

[15] VIJZELAAR S,BOS H,FOKKINK W. Brief announcement:a shared disk on distributed storage[C]//Proceedings of the 29th ACM SIGACT-SIGOPS symposium on principles of distributed computing. Zürich,Switzerland:ACM,2010:79-80.

[19] SZELIGA B,NGUYEN T,SHI Weisong. DiSK:a distributed shared disk cache for HPC environments[C]//International conference on collaborative computing:networking,applications and worksharing. Washington D. C,USA:IEEE,2009:1-8.