Gfarm カーネルドライバの

高速通信機能実装作業

操作説明書

数理技研

2014 年 3 月 31 日

更新履歴

2012/03/29 初版
2012/04/05 接続キャッシュ数の説明、エラーメッセージ表示を修正
2012/04/16 カーネルモジュールのデフォルトログレベルを DEBUG から INFO に変更
2013/03/29 タイトル変更。ファイル I/O 対応について追記
2014/03/31 タイトル変更。高速通信対応について追記
2014/03/31 不要になった libevent, fuse の記事を削除
2014/03/31 READDIR 処理変更について追記

1. はじめに

本文書は、Linux カーネルモジュール版 Gfarm の操作方法を説明したものである。
Gfarm のメタサーバーなど、既存部分の操作方法については、既存のドキュメントを参照のこと。

2. 動作確認環境

本システムは、以下の環境で動作確認した。それ以外の環境(別ディストリビューション、別カーネルバージョン)では、コンパイルエラーや、予期しない振る舞いとなる可能性がある。

項目	値	確認方法
CPU	Intel 64bit CPU (x86_64)	`uname -m`
OS	CentOS Linux release 6.4 (Final)	`cat /etc/redhat-release`
kernel	2.6.32-358.el6.x86_64	`uname -r`
gcc	4.4.6 20110731	`gcc -v`

なお、gfarm2 カーネルモジュールは、sunrpc カーネルモジュールを利用する。
/lib/modules/`uname -r`/kernel/net/sunrpc/sunrpc.ko があるか、これらがカーネルに組み込まれている必要がある。

3. セットアップ

本バージョンでは READDIR プロトコルが変更されているので、新しいクライアントを使う場合には先に gfmd を新しくしておく必要がある。

3.1 コンパイル環境の準備

3.1.1 Linux カーネルソース

カーネルモジュールのコンパイルには、カーネルソース(ただし、ヘッダファイルなどのみでよい)が必要である。

root で以下を実行する。

```
# yum install kernel-devel
```

この後、/usr/src/kernels/`uname -r` (つまり/usr/src/kernels/2.6.32-358.el6.x86_64)に、Makefile やヘッダファイルなどがあることを確認する。.c ファイルは無くて良い。

これ以外の場所にカーネルソースをおく場合は、linux/kernel/Malefile の REAL_ROOTDIR にそのディレクトリを設定すること。

3.1.2 InfiniBand モジュール

InfiniBand による高速機能を使うためには、動作マシンに InfinBand が装着されていて、
且つ、ネットワーク内の他クライアントと InfiniBand で接続されている必要がある。
InfiniBand モジュールの API を利用するので、このモジュールの情報を確かめる。

```
% modinfo ib_core
filename:/lib/modules/2.6.32.358.el6.x86_64/kernel/drivers/infiniband/core/ib
_core.ko
```

上記のように標準ドライバであれば問題ないが、extra や updates のモジュールである場合は、
当該モジュールの提供元からヘッダファイルなどの開発環境をインストールしておく必要がある。
ソースディレクトリで以下のコマンドで configure のオプションが表示されたらこれを利用してもよい。

```
% ./linux/config/findIBmodules
Configure with following options.
--with-ibsymvers=/usr/src/ofa_kernel/default/Module.symvers
--with-ib-include=/usr/src/ofa_kernel/default/include
```

configure で指定のない場合は、configure の中で上記探索が行われる。

3.2 ビルド

適当なディレクトリに納品ソース一式を置き、以下を実行する。

```
% ./configure --enable-linuxkernel
% make
```

“--enable-linuxkernel”が付くことで、トップディレクトリ直下の linux ディレクトリでも make が行われる（それだけの違いしかない）。

3.3 インストール

その後、root で以下を実行する。

```
# make install
```

通常の Gfarm インストールに加え、以下のファイルが出来たことを確認すること。

- /sbin/mount.gfarm (mount コマンドの仕様により、常に/sbin に置く)
- /usr/local/bin/ug_idmapd (インストールパスは、./configure の prefix 指定に従う)
- /lib/modules/`uname -r`/extra/gfarm2.ko

- /etc/init.d/gfsk

3.4 簡易起動確認

1. カーネルモジュールがロード可能か(root で実行)

```
# /sbin/modprobe gfarm2
# /sbin/lsmod | grep gfarm2
gfarm2                316606  2
sunrpc                241630  2 gfarm2
# /sbin/rmmod gfarm2
```

2. ug_idmapd が起動可能か(root で実行)

```
# /etc/init.d/gfsk start
# ps ax | grep ug_idmapd
1601 ?        Ss     0:00 /usr/local/bin/ug_idmapd
  -P /var/run/ug_idmapd.pid

# /etc/init.d/gfsk stop
```

4. ユーザー・グループ情報の整備

4.1 目的

カーネルモジュール版では、我々のファイルシステムモジュールに処理が渡る前に、ユーザー・グループアクセス権限のチェックがカーネルの VFS 層によって行われる。このため、Gfarm 世界のユーザー・グループ情報と、本システムを利用する環境のユーザー・グループ設定を同一にしておかなければ、意図しないアクセス権限エラーになってしまう。
また ug idmapd は、Gfarm 世界のユーザー・グループ名文字列を ID 番号に変換する際、該当するアカウントがなければ nobody にする。マウント後に ls –l などして、nobody が見えたら、そのシステムに該当アカウント(gfls –l で確認)を追加すること。

4.2 Gfarm 世界のユーザーアカウント

Gfarm 世界のユーザーアカウントは”gfuser –l”で参照できる。それらのユーザーによって本システムを利用するならば、そのアカウントが Linux システムにも無ければならない。
特に、Gfarm のトップディレクトリのユーザーは gfarmadm であるので、このアカウント設定の確認は重要である。

1. Gfarm 世界のユーザーアカウント確認

```
# gfuser -l
gfarmadm:Gfarm administrator:/:
gfarm:Gfarm administrator:/:など
```

2. アカウントの存在確認

```
# id gfarmadm
# id gfarm など
```

3. 無ければアカウント追加

```
# /usr/sbin/adduser gfarmadm [任意のオプション]
```

4. 他のユーザーについても、同様に確認する

4.3 Gfarm 世界のグループアカウント

ユーザーアカウント同様、グループアカウントを”gfgroup -l”で確認し、不在アカウントの作成と、グループ参加ユーザーの登録を行う。
特に、Gfarm のトップディレクトリのユーザー・グループは gfarmadm であるので、このアカウント設定の確認は重要である。

1. アカウントの存在確認(ID 番号は任意)

```
# id gfarmadm
uid=504(gfarmadm) gid=504(gfarmadm)
groups=504(gfarmadm)
```

2. Gfarm 世界の gfarmadm グループ参加ユーザー確認（参加ユーザーは任意）

```
# gfgroup -l gfarmadm
gfarmadm: gfarm gfarmadm gfarm2
```

3. Linux システムのグループ参加ユーザー確認(gfarm, gfarm2 が参加していない例)

```
# groupmems -g gfarmadm -l
gfarmadm
```

4. 欠けているユーザーをグループに追加する(gfarm, gfarm2 を追加する例)

```
# groupmems -g gfarmadm -a gfarm
# groupmems -g gfarmadm -a gfarm2
```

5. 他のグループについても、同様に確認する

4.4 Linux 世界のアカウント

本カーネルモジュール経由で Gfarm にアクセスしようとするユーザーが Gfarm 世界に登録されて

いなければ、通常の Gfarm セットアップの手順で、Gfarm 世界にそのアカウントを登録する。
例えば、Gfarm 世界に root アカウントがない場合（適切に設定されていない場合）、root 権限でカーネルモジュールからアクセスしようとしても、すべて権限エラーになる(mkdir などはもちろん、参照だけの df なども出来ない)。

5. SELinux の設定

SELinux が有効(enforcing)になっている場合、独自ファイルシステムである gfarm ファイルシステムに対するアクセスは、すべて禁止されてしまう。このため、SELinux を無効(permissive または disabled)にするか、適切に設定して gfarm ファイルシステムを許可するポリシーを設定する必要がある。

1. 現在の SELinux 設定を見る。Enforcing と表示された場合は、設定変更する必要がある。

```
# getenforce
```

2. 一時的に Permissive(警告を/var/log/messages に出すが、アクセスエラーにしない)に切り替えるには、以下のようにする（再起動すると Enforcing に戻ることに注意）

```
# setenforce Permissive
```

3. 恒久的に SELinux を無効にする(警告も出さず、アクセスエラーにもしない)には、/etc/selinux/config を以下のように修正し、再起動する。

```
SELINUX=disabled
```

Gfarm ファイルシステム用のアクセスポリシーを設定するには、以下の方法がある。詳しくは、SELinux 解説サイトを参照のこと。

1. SELinux ポリシー管理ツールをインストールする

```
# yum install policycoreutils-python
```

2. SELinux が Enforcing か Permissive の状態で、Gfarm カーネルモジュール経由でアクセスし、アクセス警告エラーを/var/log/messages に出す（以下は表示例）。

```
Mar 22 09:50:54 host kernel: type=1400 audit(1326439635.911:14):
avc:  denied  { associate } for  pid=18853 comm="mkdir" name="dir"
scontext=unconfined_u:object_r:unlabeled_t:s0
tcontext=system_u:object_r:unlabeled_t:s0 tclass=filesystem
```

3. 以下のコマンドを root で実行すると、ポリシー例が表示される。

```
# audit2allow -i /var/log/messages -m gfarm2
module gfarm2 1.0;
```

```
require {
        type unlabeled_t;
        class filesystem associate;
}

#============= unlabeled_t ==============
allow unlabeled_t self:filesystem associate;
```

4. このポリシーで良ければ、ポリシーを作成し、登録する

```
#  audit2allow -i /var/log/messages -M gfarm2
******************** IMPORTANT ************************
To make this policy package active, execute:

semodule -i gfarm2.pp

# semodule -i gfarm2.pp
```

6. 利用方法

6.1 利用前確認事項

- Gfarm カーネルモジュールのセットアップ(3.2 章参照)、アカウント設定(4 章)、SELinux 設定(6 章)が済んでいること
- その環境に適切な gfarm2.conf があり、gfls などができること

- 対象 Gfarm システムの gfsd は、起動していてもいなくても構わない。ただし起動していないとファイル I/O 処理は出来ない。

- gfmd のデフォルトのコネクションキャッシュ数は config.c にあるとおり 8 である(GFARM_GFMD_CONNECTION_CACHE_DEFAULT の定義参照)。このため 9 人目が接続するとそれまでキャッシュしていた接続が一つ消える。接続キャッシュ数を増やすには gfarm2.conf に以下の記述を追加すること。

```
gfmd_connection_cache 最大コネクション数
```

- デバッグする場合、エラーログにソースファイル名と行番号を表示するよう、gfarm2.conf に以下を追記しておくと良い。

```
log_level debug
log_message_verbose_level 99
```

6.2 カーネルモジュールロードと ug_idmapd 起動

root で以下を実行する。

```
# /etc/init.d/gfsk start
```

起動時にエラーメッセージが出ていないことのほか、以下を確認。

- モジュールがロードされたか
  - “/sbin/lsmod | grep gfarm2” で gfarm2 や、それが参照する,sunrpc モジュールが表示されるか
- ug_idmapd が起動したか
  - “ps ax | grep ug_idmapd”で確認
- ファイルの存在を確認する lookup 処理には、FSTAT を利用している。現在存在しないディレクトリを新規生成する場合、先に呼ばれる lookup 処理による FSTAT が必ず ENOENT エラーで終了する。デバッグログにそのエラーログが出ても無視して良い。

カーネルモジュールは、ロード時にオプション引数を指定できる。通常は指定不要である。

引数名	意味	デフォルト値
gflog_level	ログ出力レベル (0=LOG_EMERG, …, 7=LOG_DEBUG)	6 (LOG_INFO)
ug_timeout_sec	ug_idmapd との接続タイムアウト時間（秒）	1

例えば以下のように利用する。必要ならば/etc/init.d/gfsk スクリプトの該当行を修正する。

```
# /sbin/modprobe gfarm2 gflog_level=7 ug_timeout_sec=10
```

6.3 マウント

マウントポイントを /mnt/gfarm、利用ユーザー名を gfarm とすると、 root で以下を実行する。Gfarm ユーザーは、Gfarm システムにアクセスするための共有鍵ファイルの設定などが済んでいるものとする。

root で以下を実行する。

```
# mount -t gfarm \
     -o conf_path=/usr/local/etc/gfarm2.conf,luser=gfarm \
     /dev/gfarm /mnt/gfarm
```

マウント成功したか、以下を確認する。

- /proc/mounts に上記に該当する行があるか
- /etc/mtab に上記に該当する行があるか
- gfmd 稼働ホストの/var/log/messages に、luser で指定したユーザーの認証成功ログが出ているか

gfarm2 カーネルモジュールがロードされていない状態でマウントすると、/dev/gfarm が無いため、ENOENT (No such file or directory)エラーとなる。
モジュールロード済みでも、ug_idmapd が起動していなければ、/var/log/messages に以下の警告メッセージが出る。

```
libgfarm: [000000] ugidmap: ID mapping failed: has idmapd started?
```

マウントでは以下のオプションを指定できる。

引数名	意味	デフォルト値
conf_path=path	Gfarm2.conf のパス	/usr/local/etc/gfarm2.conf
luser=name	メタデータサーバとの接続を行うユーザーのローカルユーザ名。	起動ユーザ
key_path=path	共通鍵方式の鍵ファイルのパス	ホームディレクトリの鍵ファイル
readahead=page	先読みページサイズ	linux 標準
ra_async=off	先読みを非同期に行わない	行う(1)
ib_mtu=byte	InfiniBand の通信 mtu	装着 NIC の mtu
ib_port=no	InfiniBand の使用ポート番号	若い番号
ib_qkey=val	InfiniBand 相互通信の qkey	111
ib_num_rrpc=val	InfiniBand の同時受信要求数	128
ib_num_srpc=val	InfiniBand の同時送信要求数	64

6.4 ファイルシステム利用

Gfarm システムにアクセス可能なユーザーアカウントになって、マウントポイント以下のファイルシステムを利用する。

以下は、注意・制限事項である。

- 本システムは、通常のファイルシステムにある一部の機能をサポートしていない。特に、シンボリックリンクはサポートしていない。
- 意図しないアクセス拒否になった場合は、アカウント設定(4 章)と SELinux 設定(6 章)を参考に適切な設定を行った上で、モジュールロードからやり直すこと。
- gfsd が一つも起動していない場合、gfmd に対する STATFS 応答では、利用ブロック数・利用

可能ブロック数がゼロとなる。dfコマンドは、マウントしていても利用ブロック数がゼロの場合は表示しない。そういう場合は、”df –a”や”df マウントポイント”で表示させること。man dfには以下のようにある。

```
-a, --all
  サイズが 0 ブロックのファイルシステムやタイプが ‘ignore’ または‘auto’ のファイルシステムもリスト表示に含める (デフォルトでは 省かれる)。
```

6.5 アンマウント

rootで以下を実行する。

```
# umount /mnt/gfarm
```

6.5 カーネルモジュールのアンロードと ug_idmapd 停止

rootで以下を実行する。

```
# /etc/init.d/gfsk stop
```