Subject: [patch 00/10] mount ownership and unprivileged mount syscall (v3)
Posted by Miklos Szeredi on Mon, 16 Apr 2007 11:03:08 GMT
View Forum Message <> Reply to Message

This patchset adds support for keeping mount ownership information in
the kernel, and allow unprivileged mount(2) and umount(2) in certain
cases.

This can be useful for the following reasons:

- mount(8) can store ownership ("user=XY" option) in the kernel
  instead, or in addition to storing it in /etc/mtab. For example if
  private namespaces are used with mount propagations /etc/mtab
  becomes unworkable, but using /proc/mounts works fine

- fuse won't need a special suid-root mount/umount utility. Plain
  umount(8) can easily be made to work with unprivileged fuse mounts

- users can use bind mounts without having to pre-configure them in
  /etc/fstab

The following security measures are taken for unprivileged mounts:

 - only allow submounting under mounts which have a special mount flag set
 - only allow mounting on files/directories writable by the user
 - limit the number of user mounts
 - force "nosuid,nodev" mount options

Changes from the previous submissions:

 - add mount flags to set/clear mnt_flags individually
 - add "usermnt" mount flag. If it is set, then allow unprivileged
   submounts under this mount
 - make max number of user mounts default to 1024, since now the
   usermnt flag will prevent user mounts by default

--

_____
Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers

Subject: [patch 01/10] add user mounts to the kernel
Posted by Miklos Szeredi on Mon, 16 Apr 2007 11:03:09 GMT
View Forum Message <> Reply to Message

From: Miklos Szeredi <mszeredi@suse.cz>

Add ownership information to mounts.

A new mount flag, MS_SETUSER is used to make a mount owned by a user.
If this flag is specified, then the owner will be set to the current
real user id and the mount will be marked with the MNT_USER flag. On
remount don't preserve previous owner, and treat MS_SETUSER as for a
new mount. The MS_SETUSER flag is ignored on mount move.

The MNT_USER flag is not copied on any kind of mount cloning:
namespace creation, binding or propagation. For bind mounts the
cloned mount(s) are set to MNT_USER depending on the MS_SETUSER mount
flag. In all the other cases MNT_USER is always cleared.

For MNT_USER mounts a "user=UID" option is added to /proc/PID/mounts.
This is compatible with how mount ownership is stored in /etc/mtab.

It is expected, that in the future mount(8) will use MS_SETUSER to
store mount ownership within the kernel. This would help in
situations, where /etc/mtab is difficult or impossible to work with,
e.g. when using mount propagation.

Signed-off-by: Miklos Szeredi <mszeredi@suse.cz>

Index: linux/fs/namespace.c
===
--- linux.orig/fs/namespace.c 2007-04-13 12:24:15.000000000 +0200
+++ linux/fs/namespace.c 2007-04-13 13:03:44.000000000 +0200
@@ -227,6 +227,13 @@ static struct vfsmount *skip_mnt_tree(st
  return p;
 }

+static void set_mnt_user(struct vfsmount *mnt)
+{
+ BUG_ON(mnt->mnt_flags & MNT_USER);
+ mnt->mnt_uid = current->uid;
+ mnt->mnt_flags |= MNT_USER;
+}
+
 static struct vfsmount *clone_mnt(struct vfsmount *old, struct dentry *root,
       int flag)
 {
@@ -241,6 +248,11 @@ static struct vfsmount *clone_mnt(struct
   mnt->mnt_mountpoint = mnt->mnt_root;
   mnt->mnt_parent = mnt;

+ /* don't copy the MNT_USER flag */

```
+  mnt->mnt_flags &= ~MNT_USER;
+  if (flag & CL_SETUSER)
+   set_mnt_user(mnt);
+
   if (flag & CL_SLAVE) {
    list_add(&mnt->mnt_slave, &old->mnt_slave_list);
    mnt->mnt_master = old;
@@ -403,6 +415,8 @@ static int show_vfsmnt(struct seq_file *
   if (mnt->mnt_flags & fs_infop->flag)
    seq_puts(m, fs_infop->str);
  }
+ if (mnt->mnt_flags & MNT_USER)
+  seq_printf(m, ",user=%i", mnt->mnt_uid);
  if (mnt->mnt_sb->s_op->show_options)
   err = mnt->mnt_sb->s_op->show_options(m, mnt);
  seq_puts(m, " 0 0\n");
@@ -920,8 +934,9 @@ static int do_change_type(struct nameida
 /*
  * do loopback mount.
  */
-static int do_loopback(struct nameidata *nd, char *old_name, int recurse)
+static int do_loopback(struct nameidata *nd, char *old_name, int flags)
 {
+ int clone_flags;
  struct nameidata old_nd;
  struct vfsmount *mnt = NULL;
  int err = mount_is_safe(nd);
@@ -941,11 +956,12 @@ static int do_loopback(struct nameidata
  if (!check_mnt(nd->mnt) || !check_mnt(old_nd.mnt))
   goto out;

+ clone_flags = (flags & MS_SETUSER) ? CL_SETUSER : 0;
  err = -ENOMEM;
- if (recurse)
-  mnt = copy_tree(old_nd.mnt, old_nd.dentry, 0);
+ if (flags & MS_REC)
+  mnt = copy_tree(old_nd.mnt, old_nd.dentry, clone_flags);
  else
-  mnt = clone_mnt(old_nd.mnt, old_nd.dentry, 0);
+  mnt = clone_mnt(old_nd.mnt, old_nd.dentry, clone_flags);

  if (!mnt)
   goto out;
@@ -987,8 +1003,11 @@ static int do_remount(struct nameidata *

  down_write(&sb->s_umount);
  err = do_remount_sb(sb, flags, data, 0);
- if (!err)
```

```
+ if (!err) {
   nd->mnt->mnt_flags = mnt_flags;
+  if (flags & MS_SETUSER)
+   set_mnt_user(nd->mnt);
+ }
  up_write(&sb->s_umount);
  if (!err)
   security_sb_post_remount(nd->mnt, flags, data);
@@ -1093,10 +1112,13 @@ static int do_new_mount(struct nameidata
  if (!capable(CAP_SYS_ADMIN))
   return -EPERM;

- mnt = do_kern_mount(type, flags, name, data);
+ mnt = do_kern_mount(type, flags & ~MS_SETUSER, name, data);
  if (IS_ERR(mnt))
   return PTR_ERR(mnt);

+ if (flags & MS_SETUSER)
+  set_mnt_user(mnt);
+
  return do_add_mount(mnt, nd, mnt_flags, NULL);
 }

@@ -1127,7 +1149,8 @@ int do_add_mount(struct vfsmount *newmnt
  if (S_ISLNK(newmnt->mnt_root->d_inode->i_mode))
   goto unlock;

- newmnt->mnt_flags = mnt_flags;
+ /* MNT_USER was set earlier */
+ newmnt->mnt_flags |= mnt_flags;
  if ((err = graft_tree(newmnt, nd)))
   goto unlock;

@@ -1447,7 +1470,7 @@ long do_mount(char *dev_name, char *dir_
   retval = do_remount(&nd, flags & ~MS_REMOUNT, mnt_flags,
         data_page);
  else if (flags & MS_BIND)
-  retval = do_loopback(&nd, dev_name, flags & MS_REC);
+  retval = do_loopback(&nd, dev_name, flags);
  else if (flags & (MS_SHARED | MS_PRIVATE | MS_SLAVE | MS_UNBINDABLE))
   retval = do_change_type(&nd, flags);
  else if (flags & MS_MOVE)
Index: linux/include/linux/fs.h
===================================================================
--- linux.orig/include/linux/fs.h 2007-04-13 12:24:15.000000000 +0200
+++ linux/include/linux/fs.h 2007-04-13 13:03:42.000000000 +0200
@@ -123,6 +123,7 @@ extern int dir_notify_enable;
 #define MS_SLAVE (1<<19) /* change to slave */
```

```
 #define MS_SHARED (1<<20) /* change to shared */
 #define MS_RELATIME (1<<21) /* Update atime relative to mtime/ctime. */
+#define MS_SETUSER (1<<22) /* set mnt_uid to current user */
 #define MS_ACTIVE (1<<30)
 #define MS_NOUSER (1<<31)

Index: linux/include/linux/mount.h
===================================================================
--- linux.orig/include/linux/mount.h 2007-04-13 12:24:15.000000000 +0200
+++ linux/include/linux/mount.h 2007-04-13 13:03:58.000000000 +0200
@@ -30,6 +30,7 @@ struct mnt_namespace;
 #define MNT_RELATIME 0x20

 #define MNT_SHRINKABLE 0x100
+#define MNT_USER 0x200

 #define MNT_SHARED 0x1000 /* if the vfsmount is a shared mount */
 #define MNT_UNBINDABLE 0x2000 /* if the vfsmount is a unbindable mount */
@@ -61,6 +62,8 @@ struct vfsmount {
  atomic_t mnt_count;
  int mnt_expiry_mark;  /* true if marked for expiry */
  int mnt_pinned;
+
+ uid_t mnt_uid;   /* owner of the mount */
 };

 static inline struct vfsmount *mntget(struct vfsmount *mnt)
Index: linux/fs/pnode.h
===================================================================
--- linux.orig/fs/pnode.h 2007-04-13 12:24:15.000000000 +0200
+++ linux/fs/pnode.h 2007-04-13 12:25:52.000000000 +0200
@@ -22,6 +22,7 @@
 #define CL_COPY_ALL   0x04
 #define CL_MAKE_SHARED   0x08
 #define CL_PROPAGATION   0x10
+#define CL_SETUSER  0x20

 static inline void set_mnt_shared(struct vfsmount *mnt)
 {

--

_____
Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers
```

Subject: [patch 02/10] allow unprivileged umount
Posted by Miklos Szeredi on Mon, 16 Apr 2007 11:03:10 GMT
View Forum Message <> Reply to Message

From: Miklos Szeredi <mszeredi@suse.cz>

The owner doesn't need sysadmin capabilities to call umount().

Similar behavior as umount(8) on mounts having "user=UID" option in
/etc/mtab. The difference is that umount also checks /etc/fstab,
presumably to exclude another mount on the same mountpoint.

Signed-off-by: Miklos Szeredi <mszeredi@suse.cz>

Index: linux/fs/namespace.c
===
--- linux.orig/fs/namespace.c 2007-04-11 20:07:51.000000000 +0200
+++ linux/fs/namespace.c 2007-04-11 20:08:05.000000000 +0200
@@ -659,6 +659,25 @@ static int do_umount(struct vfsmount *mn
 }

 /*
+ * umount is permitted for
+ * - sysadmin
+ * - mount owner, if not forced umount
+ */
+static bool permit_umount(struct vfsmount *mnt, int flags)
+{
+ if (capable(CAP_SYS_ADMIN))
+ return true;
+
+ if (!(mnt->mnt_flags & MNT_USER))
+ return false;
+
+ if (flags & MNT_FORCE)
+ return false;
+
+ return mnt->mnt_uid == current->uid;
+}
+
+/*
  * Now umount can handle mount points as well as block devices.
  * This is important for filesystems which use unnamed block devices.
  *
@@ -681,7 +700,7 @@ asmlinkage long sys_umount(char __user *
   goto dput_and_out;

  retval = -EPERM;

```
- if (!capable(CAP_SYS_ADMIN))
+ if (!permit_umount(nd.mnt, flags))
   goto dput_and_out;

  retval = do_umount(nd.mnt, flags);

--
```

_____

Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers

---

Subject: [patch 03/10] account user mounts
Posted by Miklos Szeredi on Mon, 16 Apr 2007 11:03:11 GMT
View Forum Message <> Reply to Message

From: Miklos Szeredi <mszeredi@suse.cz>

Add sysctl variables for accounting and limiting the number of user
mounts.

The maximum number of user mounts is set to 1024 by default. This
won't in itself enable user mounts, setting the "permit user submount"
mount flag will also be needed.

Signed-off-by: Miklos Szeredi <mszeredi@suse.cz>

```
Index: linux/include/linux/sysctl.h
===================================================================
--- linux.orig/include/linux/sysctl.h 2007-04-11 18:27:46.000000000 +0200
+++ linux/include/linux/sysctl.h 2007-04-11 20:08:16.000000000 +0200
@@ -818,6 +818,8 @@ enum
  FS_AIO_NR=18, /* current system-wide number of aio requests */
  FS_AIO_MAX_NR=19, /* system-wide maximum number of aio requests */
  FS_INOTIFY=20, /* inotify submenu */
+ FS_NR_USER_MOUNTS=21, /* int:current number of user mounts */
+ FS_MAX_USER_MOUNTS=22, /* int:maximum number of user mounts */
  FS_OCFS2=988, /* ocfs2 */
 };

Index: linux/kernel/sysctl.c
===================================================================
--- linux.orig/kernel/sysctl.c 2007-04-11 18:27:46.000000000 +0200
+++ linux/kernel/sysctl.c 2007-04-11 20:08:16.000000000 +0200
@@ -1063,6 +1063,22 @@ static ctl_table fs_table[] = {
 #endif
```

```
 #endif
  {
+  .ctl_name = FS_NR_USER_MOUNTS,
+  .procname = "nr_user_mounts",
+  .data  = &nr_user_mounts,
+  .maxlen  = sizeof(int),
+  .mode  = 0444,
+  .proc_handler = &proc_dointvec,
+ },
+ {
+  .ctl_name = FS_MAX_USER_MOUNTS,
+  .procname = "max_user_mounts",
+  .data  = &max_user_mounts,
+  .maxlen  = sizeof(int),
+  .mode  = 0644,
+  .proc_handler = &proc_dointvec,
+ },
+ {
    .ctl_name = KERN_SETUID_DUMPABLE,
    .procname = "suid_dumpable",
    .data  = &suid_dumpable,
Index: linux/Documentation/filesystems/proc.txt
===================================================================
--- linux.orig/Documentation/filesystems/proc.txt 2007-04-11 18:27:44.000000000 +0200
+++ linux/Documentation/filesystems/proc.txt 2007-04-12 13:32:14.000000000 +0200
@@ -923,6 +923,15 @@ reaches aio-max-nr then io_setup will fa
 raising aio-max-nr does not result in the pre-allocation or re-sizing
 of any kernel data structures.

+nr_user_mounts and max_user_mounts
+----------------------------------
+
+These represent the number of "user" mounts and the maximum number of
+"user" mounts respectively.  User mounts may be created by
+unprivileged users.  User mounts may also be created with sysadmin
+privileges on behalf of a user, in which case nr_user_mounts may
+exceed max_user_mounts.
+
 2.2 /proc/sys/fs/binfmt_misc - Miscellaneous binary formats
 ------------------------------------------------------------

Index: linux/fs/namespace.c
===================================================================
--- linux.orig/fs/namespace.c 2007-04-11 20:08:05.000000000 +0200
+++ linux/fs/namespace.c 2007-04-12 13:29:05.000000000 +0200
@@ -39,6 +39,9 @@ static int hash_mask __read_mostly, hash
 static struct kmem_cache *mnt_cache __read_mostly;
 static struct rw_semaphore namespace_sem;
```

```
+int nr_user_mounts;
+int max_user_mounts = 1024;
+
 /* /sys/fs */
 decl_subsys(fs, NULL, NULL);
 EXPORT_SYMBOL_GPL(fs_subsys);
@@ -227,11 +230,30 @@ static struct vfsmount *skip_mnt_tree(st
  return p;
 }

+static void dec_nr_user_mounts(void)
+{
+ spin_lock(&vfsmount_lock);
+ nr_user_mounts--;
+ spin_unlock(&vfsmount_lock);
+}
+
 static void set_mnt_user(struct vfsmount *mnt)
 {
  BUG_ON(mnt->mnt_flags & MNT_USER);
  mnt->mnt_uid = current->uid;
  mnt->mnt_flags |= MNT_USER;
+ spin_lock(&vfsmount_lock);
+ nr_user_mounts++;
+ spin_unlock(&vfsmount_lock);
+}
+
+static void clear_mnt_user(struct vfsmount *mnt)
+{
+ if (mnt->mnt_flags & MNT_USER) {
+  mnt->mnt_uid = 0;
+  mnt->mnt_flags &= ~MNT_USER;
+  dec_nr_user_mounts();
+ }
 }

 static struct vfsmount *clone_mnt(struct vfsmount *old, struct dentry *root,
@@ -283,6 +305,7 @@ static inline void __mntput(struct vfsmo
 {
  struct super_block *sb = mnt->mnt_sb;
  dput(mnt->mnt_root);
+ clear_mnt_user(mnt);
  free_vfsmnt(mnt);
  deactivate_super(sb);
 }
@@ -1023,6 +1046,7 @@ static int do_remount(struct nameidata *
  down_write(&sb->s_umount);
```

```
  err = do_remount_sb(sb, flags, data, 0);
  if (!err) {
+  clear_mnt_user(nd->mnt);
   nd->mnt->mnt_flags = mnt_flags;
   if (flags & MS_SETUSER)
    set_mnt_user(nd->mnt);
Index: linux/include/linux/fs.h
===================================================================
--- linux.orig/include/linux/fs.h 2007-04-11 20:07:51.000000000 +0200
+++ linux/include/linux/fs.h 2007-04-12 13:26:42.000000000 +0200
@@ -50,6 +50,9 @@ extern struct inodes_stat_t inodes_stat;

 extern int leases_enable, lease_break_time;

+extern int nr_user_mounts;
+extern int max_user_mounts;
+
 #ifdef CONFIG_DNOTIFY
 extern int dir_notify_enable;
 #endif

--
```

_____

Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers

---

Subject: [patch 04/10] allow per-mount flags to be set/cleared individually
Posted by Miklos Szeredi on Mon, 16 Apr 2007 11:03:12 GMT
View Forum Message <> Reply to Message

With MS_REMOUNT, it is difficult to change mount flags individually,
without touching other mount flags or options, having to parse
/proc/mounts to get the old flag values and options.

It is also difficult to change a mount flag recursively, having to
walk the mount tree in userspace.

This patch solves this problem by generalizing do_change_type() so
that not only the propagation property can be changed, but mnt_flags
can be set/cleared individually.

From: Miklos Szeredi <mszeredi@suse.cz>

Signed-off-by: Miklos Szeredi <mszeredi@suse.cz>

```
Index: linux/fs/namespace.c
===================================================================
--- linux.orig/fs/namespace.c 2007-04-13 13:04:04.000000000 +0200
+++ linux/fs/namespace.c 2007-04-13 13:04:29.000000000 +0200
@@ -955,19 +956,28 @@ out_unlock:
 /*
  * recursively change the type of the mountpoint.
  */
-static int do_change_type(struct nameidata *nd, int flag)
+static int do_change_mnt(struct nameidata *nd, int flag, int mnt_flags)
 {
  struct vfsmount *m, *mnt = nd->mnt;
  int recurse = flag & MS_REC;
- int type = flag & ~MS_REC;
+ int type = flag & (MS_SHARED | MS_PRIVATE | MS_SLAVE | MS_UNBINDABLE);

  if (nd->dentry != nd->mnt->mnt_root)
   return -EINVAL;

  down_write(&namespace_sem);
  spin_lock(&vfsmount_lock);
- for (m = mnt; m; m = (recurse ? next_mnt(m, mnt) : NULL))
-  change_mnt_propagation(m, type);
+ for (m = mnt; m; m = (recurse ? next_mnt(m, mnt) : NULL)) {
+  if (type)
+   change_mnt_propagation(m, type);
+
+  if (flag & (MS_SETFLAGS | MS_CLEARFLAGS))
+   m->mnt_flags = mnt_flags;
+  else if (flag & MS_SETFLAGS)
+   m->mnt_flags |= mnt_flags;
+  else if (flag & MS_CLEARFLAGS)
+   m->mnt_flags &= ~mnt_flags;
+ }
  spin_unlock(&vfsmount_lock);
  up_write(&namespace_sem);
  return 0;
@@ -1514,8 +1526,9 @@ long do_mount(char *dev_name, char *dir_
        data_page);
  else if (flags & MS_BIND)
   retval = do_loopback(&nd, dev_name, flags);
- else if (flags & (MS_SHARED | MS_PRIVATE | MS_SLAVE | MS_UNBINDABLE))
-  retval = do_change_type(&nd, flags);
+ else if (flags & (MS_SHARED | MS_PRIVATE | MS_SLAVE | MS_UNBINDABLE |
+     MS_SETFLAGS | MS_CLEARFLAGS))
+  retval = do_change_mnt(&nd, flags, mnt_flags);
  else if (flags & MS_MOVE)
   retval = do_move_mount(&nd, dev_name);
```

```
  else
Index: linux/include/linux/fs.h
===================================================================
--- linux.orig/include/linux/fs.h 2007-04-13 13:04:04.000000000 +0200
+++ linux/include/linux/fs.h 2007-04-13 13:04:29.000000000 +0200
@@ -127,6 +127,9 @@ extern int dir_notify_enable;
 #define MS_SHARED (1<<20) /* change to shared */
 #define MS_RELATIME (1<<21) /* Update atime relative to mtime/ctime. */
 #define MS_SETUSER (1<<22) /* set mnt_uid to current user */
+#define MS_SETFLAGS (1<<23) /* set specified mount flags */
+#define MS_CLEARFLAGS (1<<24) /* clear specified mount flags */
+/* MS_SETFLAGS | MS_CLEARFLAGS: change mount flags to specified */
 #define MS_ACTIVE (1<<30)
 #define MS_NOUSER (1<<31)


--
```

Subject: [patch 05/10] Add "permit user submounts" flag to vfsmount
Posted by Miklos Szeredi on Mon, 16 Apr 2007 11:03:13 GMT
View Forum Message <> Reply to Message

From: Miklos Szeredi <mszeredi@suse.cz>

If MNT_USERMNT flag is not set in the target vfsmount, then
unprivileged mounts will be denied.

By default this flag is cleared, and can be set on new mounts, on
remounts or with the MS_SETFLAGS option.

Signed-off-by: Miklos Szeredi <mszeredi@suse.cz>

```
Index: linux/fs/namespace.c
===================================================================
--- linux.orig/fs/namespace.c 2007-04-13 13:20:12.000000000 +0200
+++ linux/fs/namespace.c 2007-04-13 13:35:40.000000000 +0200
@@ -411,6 +411,7 @@ static int show_vfsmnt(struct seq_file *
   { MNT_NOATIME, ",noatime" },
   { MNT_NODIRATIME, ",nodiratime" },
   { MNT_RELATIME, ",relatime" },
+  { MNT_USERMNT, ",usermnt" },
   { 0, NULL }
```

```
  };
  struct proc_fs_info *fs_infop;
@@ -1505,9 +1506,11 @@ long do_mount(char *dev_name, char *dir_
   mnt_flags |= MNT_NODIRATIME;
  if (flags & MS_RELATIME)
   mnt_flags |= MNT_RELATIME;
+ if (flags & MS_USERMNT)
+  mnt_flags |= MNT_USERMNT;

  flags &= ~(MS_NOSUID | MS_NOEXEC | MS_NODEV | MS_ACTIVE |
-     MS_NOATIME | MS_NODIRATIME | MS_RELATIME);
+     MS_NOATIME | MS_NODIRATIME | MS_RELATIME | MS_USERMNT);

  /* ... and get the mountpoint */
  retval = path_lookup(dir_name, LOOKUP_FOLLOW, &nd);
Index: linux/include/linux/mount.h
===================================================================
--- linux.orig/include/linux/mount.h 2007-04-13 13:17:08.000000000 +0200
+++ linux/include/linux/mount.h 2007-04-13 13:22:17.000000000 +0200
@@ -28,6 +28,7 @@ struct mnt_namespace;
 #define MNT_NOATIME 0x08
 #define MNT_NODIRATIME 0x10
 #define MNT_RELATIME 0x20
+#define MNT_USERMNT 0x40

 #define MNT_SHRINKABLE 0x100
 #define MNT_USER 0x200
Index: linux/include/linux/fs.h
===================================================================
--- linux.orig/include/linux/fs.h 2007-04-13 13:23:05.000000000 +0200
+++ linux/include/linux/fs.h 2007-04-13 13:35:34.000000000 +0200
@@ -130,6 +130,7 @@ extern int dir_notify_enable;
 #define MS_SETFLAGS (1<<23) /* set specified mount flags */
 #define MS_CLEARFLAGS (1<<24) /* clear specified mount flags */
 /* MS_SETFLAGS | MS_CLEARFLAGS: change mount flags to specified */
+#define MS_USERMNT (1<<25) /* permit unpriv. submounts under this mount */
 #define MS_ACTIVE (1<<30)
 #define MS_NOUSER (1<<31)


--
```

Subject: [patch 06/10] propagate error values from clone_mnt
Posted by Miklos Szeredi on Mon, 16 Apr 2007 11:03:14 GMT
View Forum Message <> Reply to Message

From: Miklos Szeredi <mszeredi@suse.cz>

Allow clone_mnt() to return errors other than ENOMEM. This will be
used for returning a different error value when the number of user
mounts goes over the limit.

Fix copy_tree() to return EPERM for unbindable mounts.

Don't propagate further from dup_mnt_ns() as that copy_tree() can only
fail with -ENOMEM.

Signed-off-by: Miklos Szeredi <mszeredi@suse.cz>

Index: linux/fs/namespace.c
===
--- linux.orig/fs/namespace.c 2007-04-13 13:22:17.000000000 +0200
+++ linux/fs/namespace.c 2007-04-13 13:25:35.000000000 +0200
@@ -261,42 +261,42 @@ static struct vfsmount *clone_mnt(struct
 {
  struct super_block *sb = old->mnt_sb;
  struct vfsmount *mnt = alloc_vfsmnt(old->mnt_devname);
+ if (!mnt)
+ return ERR_PTR(-ENOMEM);

- if (mnt) {
- mnt->mnt_flags = old->mnt_flags;
- atomic_inc(&sb->s_active);
- mnt->mnt_sb = sb;
- mnt->mnt_root = dget(root);
- mnt->mnt_mountpoint = mnt->mnt_root;
- mnt->mnt_parent = mnt;
-
- /* don't copy the MNT_USER flag */
- mnt->mnt_flags &= ~MNT_USER;
- if (flag & CL_SETUSER)
-   set_mnt_user(mnt);
-
- if (flag & CL_SLAVE) {
-   list_add(&mnt->mnt_slave, &old->mnt_slave_list);
-   mnt->mnt_master = old;
-   CLEAR_MNT_SHARED(mnt);
- } else {
-   if ((flag & CL_PROPAGATION) || IS_MNT_SHARED(old))
-    list_add(&mnt->mnt_share, &old->mnt_share);

```
-   if (IS_MNT_SLAVE(old))
-    list_add(&mnt->mnt_slave, &old->mnt_slave);
-   mnt->mnt_master = old->mnt_master;
-  }
-  if (flag & CL_MAKE_SHARED)
-   set_mnt_shared(mnt);
+ mnt->mnt_flags = old->mnt_flags;
+ atomic_inc(&sb->s_active);
+ mnt->mnt_sb = sb;
+ mnt->mnt_root = dget(root);
+ mnt->mnt_mountpoint = mnt->mnt_root;
+ mnt->mnt_parent = mnt;
+
+ /* don't copy the MNT_USER flag */
+ mnt->mnt_flags &= ~MNT_USER;
+ if (flag & CL_SETUSER)
+  set_mnt_user(mnt);

-  /* stick the duplicate mount on the same expiry list
-   * as the original if that was on one */
-  if (flag & CL_EXPIRE) {
-   spin_lock(&vfsmount_lock);
-   if (!list_empty(&old->mnt_expire))
-    list_add(&mnt->mnt_expire, &old->mnt_expire);
-   spin_unlock(&vfsmount_lock);
-  }
+ if (flag & CL_SLAVE) {
+  list_add(&mnt->mnt_slave, &old->mnt_slave_list);
+  mnt->mnt_master = old;
+  CLEAR_MNT_SHARED(mnt);
+ } else {
+  if ((flag & CL_PROPAGATION) || IS_MNT_SHARED(old))
+   list_add(&mnt->mnt_share, &old->mnt_share);
+  if (IS_MNT_SLAVE(old))
+   list_add(&mnt->mnt_slave, &old->mnt_slave);
+  mnt->mnt_master = old->mnt_master;
+ }
+ if (flag & CL_MAKE_SHARED)
+  set_mnt_shared(mnt);
+
+ /* stick the duplicate mount on the same expiry list
+  * as the original if that was on one */
+ if (flag & CL_EXPIRE) {
+  spin_lock(&vfsmount_lock);
+  if (!list_empty(&old->mnt_expire))
+   list_add(&mnt->mnt_expire, &old->mnt_expire);
+  spin_unlock(&vfsmount_lock);
  }
```

```
  return mnt;
 }
@@ -782,11 +782,11 @@ struct vfsmount *copy_tree(struct vfsmou
  struct nameidata nd;

  if (!(flag & CL_COPY_ALL) && IS_MNT_UNBINDABLE(mnt))
-  return NULL;
+  return ERR_PTR(-EPERM);

  res = q = clone_mnt(mnt, dentry, flag);
- if (!q)
-  goto Enomem;
+ if (IS_ERR(q))
+  goto error;
  q->mnt_mountpoint = mnt->mnt_mountpoint;

  p = mnt;
@@ -807,8 +807,8 @@ struct vfsmount *copy_tree(struct vfsmou
    nd.mnt = q;
    nd.dentry = p->mnt_mountpoint;
    q = clone_mnt(p, p->mnt_root, flag);
-   if (!q)
-    goto Enomem;
+   if (IS_ERR(q))
+    goto error;
    spin_lock(&vfsmount_lock);
    list_add_tail(&q->mnt_list, &res->mnt_list);
    attach_mnt(q, &nd);
@@ -816,7 +816,7 @@ struct vfsmount *copy_tree(struct vfsmou
   }
  }
  return res;
-Enomem:
+ error:
  if (res) {
   LIST_HEAD(umount_list);
   spin_lock(&vfsmount_lock);
@@ -824,7 +824,7 @@ Enomem:
   spin_unlock(&vfsmount_lock);
   release_mounts(&umount_list);
  }
- return NULL;
+ return q;
 }

 /*
@@ -1009,13 +1009,13 @@ static int do_loopback(struct nameidata
   goto out;
```

```
  clone_flags = (flags & MS_SETUSER) ? CL_SETUSER : 0;
- err = -ENOMEM;
  if (flags & MS_REC)
   mnt = copy_tree(old_nd.mnt, old_nd.dentry, clone_flags);
  else
   mnt = clone_mnt(old_nd.mnt, old_nd.dentry, clone_flags);

- if (!mnt)
+ err = PTR_ERR(mnt);
+ if (IS_ERR(mnt))
   goto out;

  err = graft_tree(mnt, nd);
@@ -1563,7 +1563,7 @@ static struct mnt_namespace *dup_mnt_ns(
  /* First pass: copy the tree topology */
  new_ns->root = copy_tree(mnt_ns->root, mnt_ns->root->mnt_root,
      CL_COPY_ALL | CL_EXPIRE);
- if (!new_ns->root) {
+ if (IS_ERR(new_ns->root)) {
   up_write(&namespace_sem);
   kfree(new_ns);
   return NULL;
Index: linux/fs/pnode.c
===================================================================
--- linux.orig/fs/pnode.c 2007-04-13 12:31:04.000000000 +0200
+++ linux/fs/pnode.c 2007-04-13 13:25:35.000000000 +0200
@@ -187,8 +187,9 @@ int propagate_mnt(struct vfsmount *dest_

   source =  get_source(m, prev_dest_mnt, prev_src_mnt, &type);

-  if (!(child = copy_tree(source, source->mnt_root, type))) {
-   ret = -ENOMEM;
+  child = copy_tree(source, source->mnt_root, type);
+  if (IS_ERR(child)) {
+   ret = PTR_ERR(child);
    list_splice(tree_list, tmp_list.prev);
    goto out;
   }

--
```

Subject: [patch 07/10] allow unprivileged bind mounts
Posted by Miklos Szeredi on Mon, 16 Apr 2007 11:03:15 GMT
View Forum Message <> Reply to Message

From: Miklos Szeredi <mszeredi@suse.cz>

Allow bind mounts to unprivileged users if the following conditions
are met:

- user submounts are permitted on the mountpoint's mount
- mountpoint is not a symlink or special file
- mountpoint is not a sticky directory or is owned by the current user
- mountpoint is writable by user
- the number of user mounts is below the maximum

Unprivileged mounts imply MS_SETUSER, and will also have the "nosuid"
and "nodev" mount flags set.

Signed-off-by: Miklos Szeredi <mszeredi@suse.cz>

Index: linux/fs/namespace.c
===
--- linux.orig/fs/namespace.c 2007-04-13 13:35:53.000000000 +0200
+++ linux/fs/namespace.c 2007-04-13 14:17:39.000000000 +0200
@@ -237,11 +237,30 @@ static void dec_nr_user_mounts(void)
 spin_unlock(&vfsmount_lock);
}

-static void set_mnt_user(struct vfsmount *mnt)
+static int reserve_user_mount(void)
+{
+ int err = 0;
+ spin_lock(&vfsmount_lock);
+ if (nr_user_mounts >= max_user_mounts && !capable(CAP_SYS_ADMIN))
+ err = -EPERM;
+ else
+ nr_user_mounts++;
+ spin_unlock(&vfsmount_lock);
+ return err;
+}
+
+static void __set_mnt_user(struct vfsmount *mnt)
{
 BUG_ON(mnt->mnt_flags & MNT_USER);
 mnt->mnt_uid = current->uid;
 mnt->mnt_flags |= MNT_USER;
+ if (!capable(CAP_SYS_ADMIN))
+ mnt->mnt_flags |= MNT_NOSUID | MNT_NODEV;

```
+}
+
+static void set_mnt_user(struct vfsmount *mnt)
+{
+ __set_mnt_user(mnt);
  spin_lock(&vfsmount_lock);
  nr_user_mounts++;
  spin_unlock(&vfsmount_lock);
@@ -260,9 +279,16 @@ static struct vfsmount *clone_mnt(struct
       int flag)
 {
  struct super_block *sb = old->mnt_sb;
- struct vfsmount *mnt = alloc_vfsmnt(old->mnt_devname);
+ struct vfsmount *mnt;
+
+ if (flag & CL_SETUSER) {
+  int err = reserve_user_mount();
+  if (err)
+   return ERR_PTR(err);
+ }
+ mnt = alloc_vfsmnt(old->mnt_devname);
  if (!mnt)
-  return ERR_PTR(-ENOMEM);
+  goto alloc_failed;

  mnt->mnt_flags = old->mnt_flags;
  atomic_inc(&sb->s_active);
@@ -274,7 +300,7 @@ static struct vfsmount *clone_mnt(struct
  /* don't copy the MNT_USER flag */
  mnt->mnt_flags &= ~MNT_USER;
  if (flag & CL_SETUSER)
-  set_mnt_user(mnt);
+  __set_mnt_user(mnt);

  if (flag & CL_SLAVE) {
   list_add(&mnt->mnt_slave, &old->mnt_slave_list);
@@ -299,6 +325,11 @@ static struct vfsmount *clone_mnt(struct
   spin_unlock(&vfsmount_lock);
  }
  return mnt;
+
+ alloc_failed:
+ if (flag & CL_SETUSER)
+  dec_nr_user_mounts();
+ return ERR_PTR(-ENOMEM);
 }

 static inline void __mntput(struct vfsmount *mnt)
```

```
@@ -746,22 +777,35 @@ asmlinkage long sys_oldumount(char __use

 #endif

-static int mount_is_safe(struct nameidata *nd)
+/*
+ * Conditions for unprivileged mounts are:
+ * - user submounts are permitted under this mount
+ * - mountpoint is not a symlink or special file
+ * - mountpoint is "absolutely" writable by user
+ *  o if it's a sticky directory, it must be owned by the user
+ *  o it must not be an append-only file/directory
+ */
+static int mount_is_safe(struct nameidata *nd, int *flags)
 {
+ struct inode *inode = nd->dentry->d_inode;
+
   if (capable(CAP_SYS_ADMIN))
     return 0;
- return -EPERM;
-#ifdef notyet
- if (S_ISLNK(nd->dentry->d_inode->i_mode))
+
+ if (!(nd->mnt->mnt_flags & MNT_USERMNT))
     return -EPERM;
- if (nd->dentry->d_inode->i_mode & S_ISVTX) {
-  if (current->uid != nd->dentry->d_inode->i_uid)
-   return -EPERM;
- }
- if (vfs_permission(nd, MAY_WRITE))
+
+ if (!S_ISDIR(inode->i_mode) && !S_ISREG(inode->i_mode))
+  return -EPERM;
+
+ if ((inode->i_mode & S_ISVTX) && current->fsuid != inode->i_uid)
     return -EPERM;
+
+ if (vfs_permission(nd, MAY_WRITE) || IS_APPEND(inode))
+  return -EPERM;
+
+ *flags |= MS_SETUSER;
   return 0;
-#endif
 }

 static int lives_below_in_same_fs(struct dentry *d, struct dentry *dentry)
@@ -991,7 +1035,7 @@ static int do_loopback(struct nameidata
   int clone_flags;
```

```
  struct nameidata old_nd;
  struct vfsmount *mnt = NULL;
- int err = mount_is_safe(nd);
+ int err = mount_is_safe(nd, &flags);
  if (err)
   return err;
  if (!old_name || !*old_name)

--
```

_____

Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers

---

Subject: [patch 08/10] put declaration of put_filesystem() in fs.h
Posted by Miklos Szeredi on Mon, 16 Apr 2007 11:03:16 GMT
View Forum Message <> Reply to Message

From: Miklos Szeredi <mszeredi@suse.cz>

Declarations go into headers.

Signed-off-by: Miklos Szeredi <mszeredi@suse.cz>

Index: linux/fs/super.c
```
===================================================================
--- linux.orig/fs/super.c 2007-04-13 12:26:11.000000000 +0200
+++ linux/fs/super.c 2007-04-13 13:25:40.000000000 +0200
@@ -40,10 +40,6 @@
 #include <asm/uaccess.h>


-void get_filesystem(struct file_system_type *fs);
-void put_filesystem(struct file_system_type *fs);
-struct file_system_type *get_fs_type(const char *name);
-
 LIST_HEAD(super_blocks);
 DEFINE_SPINLOCK(sb_lock);

Index: linux/include/linux/fs.h
===================================================================
--- linux.orig/include/linux/fs.h 2007-04-13 13:23:22.000000000 +0200
+++ linux/include/linux/fs.h 2007-04-13 13:25:40.000000000 +0200
@@ -1922,6 +1922,8 @@ extern int vfs_fstat(unsigned int, struc

 extern int vfs_ioctl(struct file *, unsigned int, unsigned int, unsigned long);
```

```
+extern void get_filesystem(struct file_system_type *fs);
+extern void put_filesystem(struct file_system_type *fs);
 extern struct file_system_type *get_fs_type(const char *name);
 extern struct super_block *get_super(struct block_device *);
 extern struct super_block *user_get_super(dev_t);

--
```

Subject: [patch 09/10] allow unprivileged mounts
Posted by Miklos Szeredi on Mon, 16 Apr 2007 11:03:17 GMT
View Forum Message <> Reply to Message

From: Miklos Szeredi <mszeredi@suse.cz>

Define a new fs flag FS_SAFE, which denotes, that unprivileged
mounting of this filesystem may not constitute a security problem.

Since most filesystems haven't been designed with unprivileged
mounting in mind, a thorough audit is needed before setting this flag.

Signed-off-by: Miklos Szeredi <mszeredi@suse.cz>

```
Index: linux/fs/namespace.c
===================================================================
--- linux.orig/fs/namespace.c 2007-04-13 13:35:29.000000000 +0200
+++ linux/fs/namespace.c 2007-04-13 13:35:30.000000000 +0200
@@ -785,7 +785,8 @@ asmlinkage long sys_oldumount(char __use
  *  o if it's a sticky directory, it must be owned by the user
  *  o it must not be an append-only file/directory
  */
-static int mount_is_safe(struct nameidata *nd, int *flags)
+static int mount_is_safe(struct nameidata *nd, struct file_system_type *type,
+    int *flags)
 {
  struct inode *inode = nd->dentry->d_inode;

@@ -795,6 +796,9 @@ static int mount_is_safe(struct nameidat
  if (!(nd->mnt->mnt_flags & MNT_USERMNT))
   return -EPERM;

+ if (type && !(type->fs_flags & FS_SAFE))
```

```
+  return -EPERM;
+
  if (!S_ISDIR(inode->i_mode) && !S_ISREG(inode->i_mode))
   return -EPERM;

@@ -1035,7 +1039,7 @@ static int do_loopback(struct nameidata
  int clone_flags;
  struct nameidata old_nd;
  struct vfsmount *mnt = NULL;
- int err = mount_is_safe(nd, &flags);
+ int err = mount_is_safe(nd, NULL, &flags);
  if (err)
   return err;
  if (!old_name || !*old_name)
@@ -1197,26 +1201,46 @@ out:
  * create a new mount for userspace and request it to be added into the
  * namespace's tree
  */
-static int do_new_mount(struct nameidata *nd, char *type, int flags,
+static int do_new_mount(struct nameidata *nd, char *fstype, int flags,
    int mnt_flags, char *name, void *data)
 {
+ int err;
  struct vfsmount *mnt;
+ struct file_system_type *type;

- if (!type || !memchr(type, 0, PAGE_SIZE))
+ if (!fstype || !memchr(fstype, 0, PAGE_SIZE))
   return -EINVAL;

- /* we need capabilities... */
- if (!capable(CAP_SYS_ADMIN))
-  return -EPERM;
+ type = get_fs_type(fstype);
+ if (!type)
+  return -ENODEV;

- mnt = do_kern_mount(type, flags & ~MS_SETUSER, name, data);
- if (IS_ERR(mnt))
+ err = mount_is_safe(nd, type, &flags);
+ if (err)
+  goto out_put_filesystem;
+
+ if (flags & MS_SETUSER) {
+  err = reserve_user_mount();
+  if (err)
+   goto out_put_filesystem;
+ }
```

```
+
+ mnt = vfs_kern_mount(type, flags & ~MS_SETUSER, name, data);
+ put_filesystem(type);
+ if (IS_ERR(mnt)) {
+  if (flags & MS_SETUSER)
+   dec_nr_user_mounts();
   return PTR_ERR(mnt);
+ }

  if (flags & MS_SETUSER)
-  set_mnt_user(mnt);
+  __set_mnt_user(mnt);

  return do_add_mount(mnt, nd, mnt_flags, NULL);
+
+ out_put_filesystem:
+ put_filesystem(type);
+ return err;
 }

 /*
@@ -1246,7 +1270,7 @@ int do_add_mount(struct vfsmount *newmnt
  if (S_ISLNK(newmnt->mnt_root->d_inode->i_mode))
   goto unlock;

- /* MNT_USER was set earlier */
+ /* some flags may have been set earlier */
  newmnt->mnt_flags |= mnt_flags;
  if ((err = graft_tree(newmnt, nd)))
   goto unlock;
Index: linux/include/linux/fs.h
===================================================================
--- linux.orig/include/linux/fs.h 2007-04-13 13:35:29.000000000 +0200
+++ linux/include/linux/fs.h 2007-04-13 13:35:30.000000000 +0200
@@ -96,6 +96,7 @@ extern int dir_notify_enable;
 #define FS_REQUIRES_DEV 1
 #define FS_BINARY_MOUNTDATA 2
 #define FS_HAS_SUBTYPE 4
+#define FS_SAFE 8  /* Safe to mount by unprivileged users */
 #define FS_REVAL_DOT 16384 /* Check the paths ".", ".." for staleness */
 #define FS_RENAME_DOES_D_MOVE 32768 /* FS will handle d_move()
       * during rename() internally.

--
```

Subject: [patch 10/10] allow unprivileged fuse mounts
Posted by Miklos Szeredi on Mon, 16 Apr 2007 11:03:18 GMT
View Forum Message <> Reply to Message

From: Miklos Szeredi <mszeredi@suse.cz>

Use FS_SAFE for "fuse" fs type, but not for "fuseblk".

FUSE was designed from the beginning to be safe for unprivileged users. This has also been verified in practice over many years. In addition unprivileged fuse mounts require the "usermnt" mount option to be set on the parent mount, which is more strict than the current userspace policy.

This will enable future installations to remove the suid-root fusermount utility.

Don't require the "user_id=" and "group_id=" options for unprivileged mounts, but if they are present, verify them for sanity.

Disallow the "allow_other" option for unprivileged mounts.

Signed-off-by: Miklos Szeredi <mszeredi@suse.cz>

Index: linux/fs/fuse/inode.c
===
--- linux.orig/fs/fuse/inode.c 2007-04-13 14:20:23.000000000 +0200
+++ linux/fs/fuse/inode.c 2007-04-13 14:20:27.000000000 +0200
@@ -311,6 +311,19 @@ static int parse_fuse_opt(char *opt, str
  d->max_read = ~0;
  d->blksize = 512;

+ /*
+ * For unprivileged mounts use current uid/gid. Still allow
+ * "user_id" and "group_id" options for compatibility, but
+ * only if they match these values.
+ */
+ if (!capable(CAP_SYS_ADMIN)) {
+ d->user_id = current->uid;
+ d->user_id_present = 1;
+ d->group_id = current->gid;
+ d->group_id_present = 1;
+
+ }
+
  while ((p = strsep(&opt, ",")) != NULL) {
   int token;
   int value;

```
@@ -339,6 +352,8 @@ static int parse_fuse_opt(char *opt, str
   case OPT_USER_ID:
    if (match_int(&args[0], &value))
     return 0;
+   if (d->user_id_present && d->user_id != value)
+    return 0;
    d->user_id = value;
    d->user_id_present = 1;
    break;
@@ -346,6 +361,8 @@ static int parse_fuse_opt(char *opt, str
   case OPT_GROUP_ID:
    if (match_int(&args[0], &value))
     return 0;
+   if (d->group_id_present && d->group_id != value)
+    return 0;
    d->group_id = value;
    d->group_id_present = 1;
    break;
@@ -536,6 +553,10 @@ static int fuse_fill_super(struct super_
  if (!parse_fuse_opt((char *) data, &d, is_bdev))
   return -EINVAL;

+ /* This is a privileged option */
+ if ((d.flags & FUSE_ALLOW_OTHER) && !capable(CAP_SYS_ADMIN))
+  return -EPERM;
+
  if (is_bdev) {
 #ifdef CONFIG_BLOCK
   if (!sb_set_blocksize(sb, d.blksize))
@@ -639,6 +660,7 @@ static struct file_system_type fuse_fs_t
  .fs_flags = FS_HAS_SUBTYPE,
  .get_sb  = fuse_get_sb,
  .kill_sb = kill_anon_super,
+ .fs_flags = FS_SAFE,
 };

 #ifdef CONFIG_BLOCK

--
```


Subject: Re: [patch 05/10] Add "permit user submounts" flag to vfsmount
Posted by serue on Mon, 16 Apr 2007 19:20:50 GMT

Quoting Miklos Szeredi (miklos@szeredi.hu):
> From: Miklos Szeredi <mszeredi@suse.cz>
>
> If MNT_USERMNT flag is not set in the target vfsmount, then

MNT_USER and MNT_USERMNT? I claim no way will people keep those
straight. How about MNT_ALLOWUSER and MNT_USER?

-serge

> unprivileged mounts will be denied.
>
> By default this flag is cleared, and can be set on new mounts, on
> remounts or with the MS_SETFLAGS option.
>
> Signed-off-by: Miklos Szeredi <mszeredi@suse.cz>
> ---
>
> Index: linux/fs/namespace.c
> ===
> --- linux.orig/fs/namespace.c 2007-04-13 13:20:12.000000000 +0200
> +++ linux/fs/namespace.c 2007-04-13 13:35:40.000000000 +0200
> @@ -411,6 +411,7 @@ static int show_vfsmnt(struct seq_file *
> { MNT_NOATIME, ",noatime" },
> { MNT_NODIRATIME, ",nodiratime" },
> { MNT_RELATIME, ",relatime" },
> + { MNT_USERMNT, ",usermnt" },
> { 0, NULL }
> };
> struct proc_fs_info *fs_infop;
> @@ -1505,9 +1506,11 @@ long do_mount(char *dev_name, char *dir_
> mnt_flags |= MNT_NODIRATIME;
> if (flags & MS_RELATIME)
> mnt_flags |= MNT_RELATIME;
> + if (flags & MS_USERMNT)
> + mnt_flags |= MNT_USERMNT;
>
> flags &= ~(MS_NOSUID | MS_NOEXEC | MS_NODEV | MS_ACTIVE |
> - MS_NOATIME | MS_NODIRATIME | MS_RELATIME);
> + MS_NOATIME | MS_NODIRATIME | MS_RELATIME | MS_USERMNT);
>
> /* ... and get the mountpoint */
> retval = path_lookup(dir_name, LOOKUP_FOLLOW, &nd);
> Index: linux/include/linux/mount.h
> ===
> --- linux.orig/include/linux/mount.h 2007-04-13 13:17:08.000000000 +0200
> +++ linux/include/linux/mount.h 2007-04-13 13:22:17.000000000 +0200

> @@ -28,6 +28,7 @@ struct mnt_namespace;
> #define MNT_NOATIME 0x08
> #define MNT_NODIRATIME 0x10
> #define MNT_RELATIME 0x20
> +#define MNT_USERMNT 0x40
>
> #define MNT_SHRINKABLE 0x100
> #define MNT_USER 0x200
> Index: linux/include/linux/fs.h
> ===
> --- linux.orig/include/linux/fs.h 2007-04-13 13:23:05.000000000 +0200
> +++ linux/include/linux/fs.h 2007-04-13 13:35:34.000000000 +0200
> @@ -130,6 +130,7 @@ extern int dir_notify_enable;
> #define MS_SETFLAGS (1<<23) /* set specified mount flags */
> #define MS_CLEARFLAGS (1<<24) /* clear specified mount flags */
> /* MS_SETFLAGS | MS_CLEARFLAGS: change mount flags to specified */
> +#define MS_USERMNT (1<<25) /* permit unpriv. submounts under this mount */
> #define MS_ACTIVE (1<<30)
> #define MS_NOUSER (1<<31)
>
>
> --

_____

Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers

---

Subject: Re: [patch 02/10] allow unprivileged umount
Posted by ebiederm on Mon, 16 Apr 2007 19:39:19 GMT
View Forum Message <> Reply to Message

Miklos Szeredi <miklos@szeredi.hu> writes:

> From: Miklos Szeredi <mszeredi@suse.cz>
>
> The owner doesn't need sysadmin capabilities to call umount().
>
> Similar behavior as umount(8) on mounts having "user=UID" option in
> /etc/mtab. The difference is that umount also checks /etc/fstab,
> presumably to exclude another mount on the same mountpoint.
>

bool in the kernel?

int would be much more recognizable as this is not C++

Or do you have place to convert the rest of the kernel that is using

int to return a true/false value to bool?

> +static bool permit_umount(struct vfsmount *mnt, int flags)
> +{
> + if (capable(CAP_SYS_ADMIN))
> + return true;
> +
> + if (!(mnt->mnt_flags & MNT_USER))
> + return false;
> +
> + if (flags & MNT_FORCE)
> + return false;
> +
> + return mnt->mnt_uid == current->uid;
> +}

Eric

_____
Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers

---

Subject: Re: [patch 05/10] Add "permit user submounts" flag to vfsmount
Posted by Miklos Szeredi on Tue, 17 Apr 2007 10:44:39 GMT
View Forum Message <> Reply to Message

> > From: Miklos Szeredi <mszeredi@suse.cz>
> >
> > If MNT_USERMNT flag is not set in the target vfsmount, then
>
> MNT_USER and MNT_USERMNT? I claim no way will people keep those
> straight. How about MNT_ALLOWUSER and MNT_USER?

Umm, is "allowuser" more clear than "usermnt"? What is allowed to the user? "allowusermnt" may be more descriptive, but it's a bit too long.

I don't think it matters all that much, the user will have to look up the semantics in the manpage anyway. Is "nosuid" descriptive? Not very much, but we got used to it.

Miklos

_____
Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers

Subject: Re: [patch 05/10] Add "permit user submounts" flag to vfsmount
Posted by serue on Tue, 17 Apr 2007 14:33:01 GMT
View Forum Message <> Reply to Message

Quoting Miklos Szeredi (miklos@szeredi.hu):
> > > From: Miklos Szeredi <mszeredi@suse.cz>
> > >
> > > If MNT_USERMNT flag is not set in the target vfsmount, then
> >
> > MNT_USER and MNT_USERMNT? I claim no way will people keep those
> > straight. How about MNT_ALLOWUSER and MNT_USER?
>
> Umm, is "allowuser" more clear than "usermnt"? What is allowed to the

I think so, yes. One makes it clear that we're talking about allowing
user (somethings :), one might just as well mean "this is a user mount."

> user? "allowusermnt" may be more descriptive, but it's a bit too
> long.

Yes, if it weren't too long it would by far have been my preference.
Maybe despite the length we should still go with it...

> I don't think it matters all that much, the user will have to look up
> the semantics in the manpage anyway. Is "nosuid" descriptive? Not
> very much, but we got used to it.

nosuid is quite clear. MNT_USER and MNT_USERMNT are so confusing that
in the time I go from quitting the manpage to foregrounding my editor, I
may have already forgotten which was which.

-serge

_____
Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers

Subject: Re: [patch 05/10] Add "permit user submounts" flag to vfsmount
Posted by Miklos Szeredi on Tue, 17 Apr 2007 16:08:27 GMT
View Forum Message <> Reply to Message

> > > MNT_USER and MNT_USERMNT? I claim no way will people keep those
> > > straight. How about MNT_ALLOWUSER and MNT_USER?
> >
> > Umm, is "allowuser" more clear than "usermnt"? What is allowed to the
>
> I think so, yes. One makes it clear that we're talking about allowing

> user (somethings :), one might just as well mean "this is a user mount."
>
> > user? "allowusermnt" may be more descriptive, but it's a bit too
> > long.
>
> Yes, if it weren't too long it would by far have been my preference.
> Maybe despite the length we should still go with it...
>
> > I don't think it matters all that much, the user will have to look up
> > the semantics in the manpage anyway. Is "nosuid" descriptive? Not
> > very much, but we got used to it.
>
> nosuid is quite clear.

Is it? Shouldn't these be "allowsuid", "noallowsuid", "allowexec",
"noallowexec"?

See, we mentally add the "allow" quite easily.

> MNT_USER and MNT_USERMNT are so confusing that in the time I go from
> quitting the manpage to foregrounding my editor, I may have already
> forgotten which was which.

Well, to the user they are always in the form "user=123" and
"usermnt", so they are not as easy to confuse.

But I feel a bit stupid bickering about this, because it isn't so
important. "allowuser" or "allowusermnt" are fine by me if you think
they are substantially better than "usermnt".

Miklos

_____
Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers

---

Subject: Re: [patch 05/10] Add "permit user submounts" flag to vfsmount
Posted by serue on Tue, 17 Apr 2007 16:54:34 GMT
View Forum Message <> Reply to Message

Quoting Miklos Szeredi (miklos@szeredi.hu):
> > > > MNT_USER and MNT_USERMNT? I claim no way will people keep those
> > > > straight. How about MNT_ALLOWUSER and MNT_USER?
> > >
> > > Umm, is "allowuser" more clear than "usermnt"? What is allowed to the
> >
> > I think so, yes. One makes it clear that we're talking about allowing

> > user (somethings :), one might just as well mean "this is a user mount."
> >
> > > user? "allowusermnt" may be more descriptive, but it's a bit too
> > > long.
> >
> > Yes, if it weren't too long it would by far have been my preference.
> > Maybe despite the length we should still go with it...
> >
> > > I don't think it matters all that much, the user will have to look up
> > > the semantics in the manpage anyway. Is "nosuid" descriptive? Not
> > > very much, but we got used to it.
> >
> > nosuid is quite clear.
>
> Is it? Shouldn't these be "allowsuid", "noallowsuid", "allowexec",
> "noallowexec"?
>
> See, we mentally add the "allow" quite easily.

But they aren't accompanied by a flag meaning "don't allow any
non-nosuid mounts below this point". *That* is what causes the problem
here.

> > MNT_USER and MNT_USERMNT are so confusing that in the time I go from
> > quitting the manpage to foregrounding my editor, I may have already
> > forgotten which was which.
>
> Well, to the user they are always in the form "user=123" and
> "usermnt", so they are not as easy to confuse.

It still makes the kernel code harder to read, but for the user yes that
is helpful.

> But I feel a bit stupid bickering about this, because it isn't so
> important. "allowuser" or "allowusermnt" are fine by me if you think
> they are substantially better than "usermnt".

Thanks, I really really do :)

-serge