Subject: [PATCH 2/4] sysfs: Implement sysfs manged shadow directory support.
Posted by ebiederm on Fri, 22 Jun 2007 07:35:51 GMT
View Forum Message <> Reply to Message

The problem. When implementing a network namespace I need to be able
to have multiple network devices with the same name. Currently this
is a problem for /sys/class/net/*, /sys/devices/virtual/net/*, and
potentially a few other directories of the form /sys/ ... /net/*.

What I want is for each network namespace to have it's own separate
set of directories. /sys/class/net/, /sys/devices/virtual/net,
and /sys/ ... /net/, and in each set I want to name them
/sys/class/net/, sys/devices/virtual/net/ and /sys/ ... /net/ respectively.

I looked and the VFS actually allows that. All that is needed is
for /sys/class/net to implement a follow link method to redirect
lookups to the real directory you want.

I am calling the concept of multiple directories all at the same path
in the filesystem shadow directories, the directory entry that
implements the follow_link method the shadow master, and the directories
that are the target of the follow link method shadow directories.

It turns out that just implementing a follow_link method is not
quite enough. The existince of directories of the form /sys/ ... /net/
can depend on the presence or absence of hotplug hardware, which
means I need a simple race free way to create and destroy these
directories.

To achieve a race free design all shadow directories are created
and managed by sysfs itself. The upper level code that knows what
shadow directories we need provides just two methods that enable
this:
 current_tag() - that returns a "void *" tag that identifies the context of
the current process.
 kobject_tag(kobj) - that returns a "void *" tag that identifies the context
a kobject should be in.
Everything else is left up to sysfs.

For the network namespace current_tag and kobject_tag are essentially
one line functions, and look to remain that.

The work needed in sysfs is more extensive. At each directory or
symlink creation I need to check if the shadow directory it belongs
in exists and if it does not create it. Likewise at each symlink
or directory removal I need to check if sysfs directory it is being
removed from is a shadow directory and if this is the last object
in the shadow directory and if so to remove the shadow directory

as well.

I also need a bunch of boiler plate that properly finds, creates, and
removes/frees the shadow directories.

Doing all of that in sysfs isn't bad it is just a bit tedious. Being race
free is just a manner of ensure we have the directory inode mutex when
we add or remove a shadow directory. Trying to do this race free
anywhere besides in sysfs is very nasty, and requires unhealthy
amounts of information about how sysfs is implemented.

Currently only directories which hold kobjects, and
symlinks are supported. There is not enough information
in the current file attribute interfaces to give us anything
to discriminate on which makes it useless, and there are
not potential users which makes it an uniteresting problem
to solve.

Signed-off-by: Eric W. Biederman <ebiederm@xmission.com>
---
 fs/sysfs/bin.c | 2 +-
 fs/sysfs/dir.c | 347 ++-----
 fs/sysfs/file.c | 4 +-
 fs/sysfs/group.c | 12 +-
 fs/sysfs/inode.c | 15 ++-
 fs/sysfs/symlink.c | 26 +++-
 fs/sysfs/sysfs.h | 6 +-
 include/linux/sysfs.h | 14 ++
 8 files changed, 374 insertions(+), 52 deletions(-)

diff --git a/fs/sysfs/bin.c b/fs/sysfs/bin.c
index 3c5574a..b96a893 100644
--- a/fs/sysfs/bin.c
+++ b/fs/sysfs/bin.c
@@ -248,7 +248,7 @@ int sysfs_create_bin_file(struct kobject * kobj, struct bin_attribute * attr)

 void sysfs_remove_bin_file(struct kobject * kobj, struct bin_attribute * attr)
 {
- if (sysfs_hash_and_remove(kobj->dentry, attr->attr.name) < 0) {
+ if (sysfs_hash_and_remove(kobj, kobj->dentry, attr->attr.name) < 0) {
   printk(KERN_ERR "%s: "
    "bad dentry or inode or no such file: \"%s\"\n",
    __FUNCTION__, attr->attr.name);
diff --git a/fs/sysfs/dir.c b/fs/sysfs/dir.c
index b1da4fc..043464e 100644
--- a/fs/sysfs/dir.c
+++ b/fs/sysfs/dir.c
@@ -302,7 +302,8 @@ static void sysfs_attach_dentry(struct sysfs_dirent *sd, struct dentry

```
*dentry)
  sd->s_dentry = dentry;
  spin_unlock(&sysfs_lock);

- d_rehash(dentry);
+ if (dentry->d_flags & DCACHE_UNHASHED)
+  d_rehash(dentry);
 }

 void sysfs_attach_dirent(struct sysfs_dirent *sd,
@@ -348,12 +349,17 @@ static int create_dir(struct kobject *kobj, struct dentry *parent,
  umode_t mode = S_IFDIR| S_IRWXU | S_IRUGO | S_IXUGO;
  struct dentry *dentry;
  struct inode *inode;
- struct sysfs_dirent *sd;
+ struct sysfs_dirent *sd, *parent_sd;

  mutex_lock(&parent->d_inode->i_mutex);

  /* allocate */
- dentry = lookup_one_len(name, parent, strlen(name));
+ parent_sd = sysfs_resolve_for_create(kobj, parent);
+ if (IS_ERR(parent_sd)) {
+  error = PTR_ERR(parent_sd);
+  goto out_unlock;
+ }
+ dentry = lookup_one_len(name, parent_sd->s_dentry, strlen(name));
  if (IS_ERR(dentry)) {
   error = PTR_ERR(dentry);
   goto out_unlock;
@@ -382,12 +388,12 @@ static int create_dir(struct kobject *kobj, struct dentry *parent,

  /* link in */
  error = -EEXIST;
- if (sysfs_dirent_exist(parent->d_fsdata, name))
+ if (sysfs_dirent_exist(parent_sd, name))
   goto out_iput;

  sysfs_instantiate(dentry, inode);
- inc_nlink(parent->d_inode);
- sysfs_attach_dirent(sd, parent->d_fsdata, dentry);
+ inc_nlink(parent_sd->s_dentry->d_inode);
+ sysfs_attach_dirent(sd, parent_sd, dentry);

  *p_dentry = dentry;
  error = 0;
@@ -502,9 +508,11 @@ static void remove_dir(struct dentry * d)
```

```
  mutex_unlock(&parent->d_inode->i_mutex);

+ sysfs_prune_shadow_sd(sd->s_parent);
  sysfs_drop_dentry(sd);
  sysfs_deactivate(sd);
  sysfs_put(sd);
+
 }

 void sysfs_remove_subdir(struct dentry * d)
@@ -512,6 +520,64 @@ void sysfs_remove_subdir(struct dentry * d)
  remove_dir(d);
 }

+static void sysfs_empty_dir(struct dentry *dentry,
+    struct sysfs_dirent **removed)
+{
+ struct sysfs_dirent *parent_sd;
+ struct sysfs_dirent **pos;
+
+ parent_sd = dentry->d_fsdata;
+ pos = &parent_sd->s_children;
+ while (*pos) {
+  struct sysfs_dirent *sd = *pos;
+
+  if (sd->s_type && (sd->s_type & SYSFS_NOT_PINNED)) {
+   *pos = sd->s_sibling;
+   sd->s_sibling = *removed;
+   *removed = sd;
+  } else
+   pos = &(*pos)->s_sibling;
+ }
+}
+
+static void __sysfs_remove_empty_shadow(struct dentry *shadow)
+{
+ struct sysfs_dirent *sd;
+
+ if (d_unhashed(shadow))
+  return;
+
+ sd = shadow->d_fsdata;
+ sysfs_unlink_sibling(sd);
+ simple_rmdir(shadow->d_inode, shadow);
+ d_delete(shadow);
+ sysfs_put(sd);
+}
+
```

```
+static void sysfs_remove_empty_shadow(struct dentry *shadow)
+{
+ mutex_lock(&shadow->d_inode->i_mutex);
+ __sysfs_remove_empty_shadow(shadow);
+ mutex_unlock(&shadow->d_inode->i_mutex);
+}
+
+static void sysfs_remove_shadows(struct dentry *dentry,
+    struct sysfs_dirent **removed)
+{
+ struct sysfs_dirent *parent_sd, **pos;
+
+ parent_sd = dentry->d_fsdata;
+ pos = &parent_sd->s_children;
+ while (*pos) {
+  struct sysfs_dirent *sd = *pos;
+  struct dentry *shadow;
+
+  shadow = dget(sd->s_dentry);
+  sysfs_empty_dir(shadow, removed);
+  __sysfs_remove_empty_shadow(shadow);
+  dput(shadow);
+ }
+}

 /**
  * sysfs_remove_dir - remove an object's directory.
@@ -524,10 +590,8 @@ void sysfs_remove_subdir(struct dentry * d)

 void sysfs_remove_dir(struct kobject * kobj)
 {
- struct dentry *dentry = kobj->dentry;
+ struct dentry *dentry = dget(kobj->dentry);
  struct sysfs_dirent *removed = NULL;
- struct sysfs_dirent *parent_sd;
- struct sysfs_dirent **pos;

  spin_lock(&kobj_sysfs_assoc_lock);
  kobj->dentry = NULL;
@@ -538,18 +602,10 @@ void sysfs_remove_dir(struct kobject * kobj)

  pr_debug("sysfs %s: removing dir\n",dentry->d_name.name);
  mutex_lock(&dentry->d_inode->i_mutex);
- parent_sd = dentry->d_fsdata;
- pos = &parent_sd->s_children;
- while (*pos) {
-  struct sysfs_dirent *sd = *pos;
-
```

```
-  if (sd->s_type && (sd->s_type & SYSFS_NOT_PINNED)) {
-   *pos = sd->s_sibling;
-   sd->s_sibling = removed;
-   removed = sd;
-  } else
-   pos = &(*pos)->s_sibling;
- }
+ if (dentry->d_inode->i_op == &sysfs_dir_inode_operations)
+  sysfs_empty_dir(dentry, &removed);
+ else
+  sysfs_remove_shadows(dentry, &removed);
  mutex_unlock(&dentry->d_inode->i_mutex);

  while (removed) {
@@ -586,7 +642,7 @@ int sysfs_rename_dir(struct kobject * kobj, const char *new_name)
  down_write(&sysfs_rename_sem);
  mutex_lock(&inode->i_mutex);

- parent_sd = parent->d_fsdata;
+ parent_sd = sysfs_resolve_for_create(kobj, parent);
  new_dentry = lookup_one_len(new_name, parent_sd->s_dentry, strlen(new_name));
  if (IS_ERR(new_dentry)) {
   error = PTR_ERR(new_dentry);
@@ -637,6 +693,7 @@ int sysfs_rename_dir(struct kobject * kobj, const char *new_name)
  mutex_unlock(&inode->i_mutex);
  up_write(&sysfs_rename_sem);

+ sysfs_prune_shadow_sd(parent->d_fsdata);
  dput(parent);

  return error;
@@ -644,25 +701,36 @@ int sysfs_rename_dir(struct kobject * kobj, const char *new_name)

 int sysfs_move_dir(struct kobject *kobj, struct kobject *new_parent)
 {
+ struct inode *old_parent_inode, *new_parent_inode;
  struct dentry *old_parent_dentry, *new_parent_dentry, *new_dentry;
  struct sysfs_dirent *new_parent_sd, *sd;
  int error;

- old_parent_dentry = kobj->parent ?
-  kobj->parent->dentry : sysfs_mount->mnt_sb->s_root;
+ old_parent_dentry = kobj->dentry->d_parent;
  new_parent_dentry = new_parent ?
   new_parent->dentry : sysfs_mount->mnt_sb->s_root;

- if (old_parent_dentry->d_inode == new_parent_dentry->d_inode)
+ old_parent_inode = old_parent_dentry->d_inode;
```

```
+ new_parent_inode = new_parent_dentry->d_inode;
+
+ if (old_parent_inode == new_parent_inode)
   return 0; /* nothing to move */
+
+ dget(old_parent_dentry);
 again:
- mutex_lock(&old_parent_dentry->d_inode->i_mutex);
- if (!mutex_trylock(&new_parent_dentry->d_inode->i_mutex)) {
-  mutex_unlock(&old_parent_dentry->d_inode->i_mutex);
+ mutex_lock(&old_parent_inode->i_mutex);
+ if (!mutex_trylock(&new_parent_inode->i_mutex)) {
+  mutex_unlock(&old_parent_inode->i_mutex);
   goto again;
  }

- new_parent_sd = new_parent_dentry->d_fsdata;
+ new_parent_sd = sysfs_resolve_for_create(kobj, new_parent_dentry);
+ if (IS_ERR(new_parent_sd)) {
+  error = PTR_ERR(new_parent_sd);
+  goto out;
+ }
+
+ new_parent_dentry = new_parent_sd->s_dentry;
  sd = kobj->dentry->d_fsdata;

  new_dentry = lookup_one_len(kobj->name, new_parent_dentry,
@@ -684,8 +752,11 @@ again:
  sysfs_link_sibling(sd);

 out:
- mutex_unlock(&new_parent_dentry->d_inode->i_mutex);
- mutex_unlock(&old_parent_dentry->d_inode->i_mutex);
+ mutex_unlock(&new_parent_inode->i_mutex);
+ mutex_unlock(&old_parent_inode->i_mutex);
+
+ sysfs_prune_shadow_sd(old_parent_dentry->d_fsdata);
+ dput(old_parent_dentry);

  return error;
 }
@@ -754,6 +825,11 @@ static int sysfs_readdir(struct file * filp, void * dirent, filldir_t filldir)
    i++;
    /* fallthrough */
   default:
+   /* If I am the shadow master return nothing. */
+   if ((parent_sd->s_type == SYSFS_DIR) &&
+       (dentry->d_inode->i_op != &sysfs_dir_inode_operations))
```

```
+    return 0;
+
    pos = &parent_sd->s_children;
    while (*pos != cursor)
     pos = &(*pos)->s_sibling;
@@ -838,3 +914,212 @@ const struct file_operations sysfs_dir_operations = {
  .read  = generic_read_dir,
  .readdir = sysfs_readdir,
 };
+
+static struct sysfs_dirent *find_shadow_sd(struct sysfs_dirent *parent_sd, const void *target)
+{
+ /* Find the shadow directory for the specified tag */
+ struct sysfs_dirent *sd;
+
+ for (sd = parent_sd->s_children; sd; sd = sd->s_sibling) {
+  if (sd->s_name != target)
+   continue;
+  break;
+ }
+ return sd;
+}
+
+static const void *find_shadow_tag(struct kobject *kobj)
+{
+ /* Find the tag the current kobj is cached with */
+ struct sysfs_dirent *sd;
+
+ sd = kobj->dentry->d_fsdata;
+ return sd->s_parent->s_name;
+}
+
+static struct sysfs_dirent *add_shadow_sd(struct sysfs_dirent *parent_sd, const void *tag)
+{
+ struct sysfs_dirent *sd;
+ struct dentry *dir, *shadow;
+ struct inode *inode;
+ int err;
+
+ dir = parent_sd->s_dentry;
+ inode = dir->d_inode;
+
+ err = -ENOMEM;
+ shadow = d_alloc(dir->d_parent, &dir->d_name);
+ if (!shadow)
+  goto error;
+
+ sd = sysfs_new_dirent(tag, inode->i_mode, SYSFS_SHADOW);
```

```
+ if (!sd)
+  goto error;
+
+ /* Since the shadow directory is reachable make it look
+  * like it is actually hashed.
+  */
+ shadow->d_hash.pprev = &shadow->d_hash.next;
+ shadow->d_hash.next = NULL;
+ shadow->d_flags &= ~DCACHE_UNHASHED;
+
+ sysfs_instantiate(shadow, igrab(inode));
+ inc_nlink(inode);
+ inc_nlink(dir->d_parent->d_inode);
+ sysfs_attach_dirent(sd, parent_sd, shadow);
+out:
+ return sd;
+error:
+ dput(shadow);
+ sd = ERR_PTR(err);
+ goto out;
+}
+
+void sysfs_prune_shadow_sd(struct sysfs_dirent *sd)
+{
+ /* If a shadow directory goes empty remove it. */
+ struct dentry *shadow;
+
+ if (sd->s_type != SYSFS_SHADOW)
+  return;
+
+ shadow = sd->s_dentry;
+ if (!shadow)
+  return;
+
+ if (sd->s_children)
+  return;
+
+ sysfs_remove_empty_shadow(shadow);
+}
+
+static void *sysfs_shadow_follow_link(struct dentry *dentry, struct nameidata *nd)
+{
+ struct sysfs_dirent *sd;
+ struct dentry *dest;
+
+ sd = dentry->d_fsdata;
+ dest = NULL;
+ if (sd->s_type == SYSFS_DIR) {
```

```
+  const struct shadow_dir_operations *shadow_ops;
+  struct inode *inode;
+  inode = dentry->d_inode;
+  shadow_ops = inode->i_private;
+  mutex_lock(&inode->i_mutex);
+  sd = find_shadow_sd(sd, shadow_ops->current_tag());
+  if (sd)
+   dest = sd->s_dentry;
+  dget(dest);
+  mutex_unlock(&inode->i_mutex);
+ }
+ if (!dest)
+  dest = dget(dentry);
+ dput(nd->dentry);
+ nd->dentry = dest;
+
+ return NULL;
+}
+
+static struct inode_operations sysfs_shadow_inode_operations = {
+ .lookup  = sysfs_lookup,
+ .setattr = sysfs_setattr,
+ .follow_link = sysfs_shadow_follow_link,
+};
+
+
+struct sysfs_dirent *sysfs_resolve_for_create(struct kobject *kobj,
+           struct dentry *dentry)
+{
+ const struct shadow_dir_operations *shadow_ops;
+ struct sysfs_dirent *sd, *shadow_sd;
+ struct inode *inode;
+
+ sd = dentry->d_fsdata;
+ inode = dentry->d_inode;
+ shadow_ops = inode->i_private;
+ if (inode->i_op == &sysfs_shadow_inode_operations) {
+  const void *tag;
+  if (sd->s_type == SYSFS_SHADOW)
+   sd = sd->s_parent;
+  tag = shadow_ops->kobject_tag(kobj);
+  shadow_sd = find_shadow_sd(sd, tag);
+  if (!shadow_sd)
+   shadow_sd = add_shadow_sd(sd, tag);
+  sd = shadow_sd;
+ }
+ return sd;
+}
```

```
+
+struct sysfs_dirent *sysfs_resolve_for_remove(struct kobject *kobj, struct dentry *dentry)
+{
+ /* If dentry is a shadow directory find the shadow that is
+  * stored under the same tag as kobj.  This allows removal
+  * of dirents to function properly even if the value of
+  * kobject_tag() has changed since we initially created
+  * the dirents assoctated with kobj.
+  */
+ const struct shadow_dir_operations *shadow_ops;
+ struct sysfs_dirent *sd;
+ struct inode *inode;
+
+ sd = dentry->d_fsdata;
+ inode = dentry->d_inode;
+ shadow_ops = inode->i_private;
+ if (inode->i_op == &sysfs_shadow_inode_operations) {
+  if (sd->s_type == SYSFS_SHADOW)
+   sd = sd->s_parent;
+  sd = find_shadow_sd(sd, find_shadow_tag(kobj));
+  if (!sd)
+   sd = ERR_PTR(-ENOENT);
+ }
+ return sd;
+}
+
+/**
+ * sysfs_enable_shadowing - Automatically create shadows of a directory
+ * @kobj: object to automatically shadow
+ *
+ * Once shadowing has been enabled on a directory the contents
+ * of the directory become dependent upon context.
+ *
+ * shadow_ops->current_tag() returns the context for the current
+ * process.
+ *
+ * shadow_ops->kobject_tag() returns the context that a given kobj
+ * resides in.
+ *
+ * Using those methods the sysfs code on shadowed directories
+ * carefully stores the files so that when we lookup files
+ * we get the proper answer for our context.
+ *
+ * If the context of a kobject is changed it is expected that
+ * the kobject will be renamed so the appopriate sysfs data structures
+ * can be updated.
+ */
+int sysfs_enable_shadowing(struct kobject *kobj,
```

```
+ const struct shadow_dir_operations *shadow_ops)
+{
+ struct sysfs_dirent *sd;
+ struct dentry *dentry;
+ struct inode *inode;
+ int err;
+
+ dentry = kobj->dentry;
+ inode = dentry->d_inode;
+
+ mutex_lock(&inode->i_mutex);
+ /* We can only enable shadowing on empty directories
+  * where shadowing is not already enabled.
+  */
+ sd = dentry->d_fsdata;
+ err = -EINVAL;
+ if (!sd->s_children && (sd->s_type == SYSFS_DIR) &&
+     (inode->i_op == &sysfs_dir_inode_operations)) {
+  inode->i_private = (void *)shadow_ops;
+  inode->i_op = &sysfs_shadow_inode_operations;
+  err = 0;
+ }
+ mutex_unlock(&inode->i_mutex);
+ return err;
+}
+
diff --git a/fs/sysfs/file.c b/fs/sysfs/file.c
index a84b734..8509ae1 100644
--- a/fs/sysfs/file.c
+++ b/fs/sysfs/file.c
@@ -560,7 +560,7 @@ EXPORT_SYMBOL_GPL(sysfs_chmod_file);

 void sysfs_remove_file(struct kobject * kobj, const struct attribute * attr)
 {
- sysfs_hash_and_remove(kobj->dentry, attr->name);
+ sysfs_hash_and_remove(kobj, kobj->dentry, attr->name);
 }


@@ -577,7 +577,7 @@ void sysfs_remove_file_from_group(struct kobject *kobj,

  dir = lookup_one_len(group, kobj->dentry, strlen(group));
  if (!IS_ERR(dir)) {
-  sysfs_hash_and_remove(dir, attr->name);
+  sysfs_hash_and_remove(kobj, dir, attr->name);
   dput(dir);
  }
 }
```

```
diff --git a/fs/sysfs/group.c b/fs/sysfs/group.c
index 82e0f55..0ae49ab 100644
--- a/fs/sysfs/group.c
+++ b/fs/sysfs/group.c
@@ -17,16 +17,16 @@
 #include "sysfs.h"


-static void remove_files(struct dentry * dir,
+static void remove_files(struct kobject *kobj, struct dentry * dir,
     const struct attribute_group * grp)
 {
  struct attribute *const* attr;

  for (attr = grp->attrs; *attr; attr++)
-  sysfs_hash_and_remove(dir,(*attr)->name);
+  sysfs_hash_and_remove(kobj, dir,(*attr)->name);
 }

-static int create_files(struct dentry * dir,
+static int create_files(struct kobject *kobj, struct dentry * dir,
    const struct attribute_group * grp)
 {
  struct attribute *const* attr;
@@ -36,7 +36,7 @@ static int create_files(struct dentry * dir,
   error = sysfs_add_file(dir, *attr, SYSFS_KOBJ_ATTR);
  }
  if (error)
-  remove_files(dir,grp);
+  remove_files(kobj, dir, grp);
  return error;
 }

@@ -56,7 +56,7 @@ int sysfs_create_group(struct kobject * kobj,
  } else
   dir = kobj->dentry;
  dir = dget(dir);
- if ((error = create_files(dir,grp))) {
+ if ((error = create_files(kobj, dir, grp))) {
   if (grp->name)
    sysfs_remove_subdir(dir);
  }
@@ -77,7 +77,7 @@ void sysfs_remove_group(struct kobject * kobj,
  else
   dir = dget(kobj->dentry);

- remove_files(dir,grp);
+ remove_files(kobj, dir, grp);
```

```
  if (grp->name)
   sysfs_remove_subdir(dir);
  /* release the ref. taken in this routine */
diff --git a/fs/sysfs/inode.c b/fs/sysfs/inode.c
index 8181c49..2ac07fd 100644
--- a/fs/sysfs/inode.c
+++ b/fs/sysfs/inode.c
@@ -272,10 +272,11 @@ void sysfs_drop_dentry(struct sysfs_dirent *sd)
   dput(parent);
 }

-int sysfs_hash_and_remove(struct dentry * dir, const char * name)
+int sysfs_hash_and_remove(struct kobject *kobj, struct dentry * dir, const char * name)
 {
+ struct inode * inode;
  struct sysfs_dirent **pos, *sd;
- struct sysfs_dirent *parent_sd = dir->d_fsdata;
+ struct sysfs_dirent *parent_sd;
  int found = 0;

  if (!dir)
@@ -285,7 +286,11 @@ int sysfs_hash_and_remove(struct dentry * dir, const char * name)
   /* no inode means this hasn't been made visible yet */
   return -ENOENT;

- mutex_lock_nested(&dir->d_inode->i_mutex, I_MUTEX_PARENT);
+ inode = dir->d_inode;
+ mutex_lock_nested(&inode->i_mutex, I_MUTEX_PARENT);
+ parent_sd = sysfs_resolve_for_remove(kobj, dir);
+ if (IS_ERR(parent_sd))
+  goto out_unlock;
  for (pos = &parent_sd->s_children; *pos; pos = &(*pos)->s_sibling) {
   sd = *pos;

@@ -298,11 +303,13 @@ int sysfs_hash_and_remove(struct dentry * dir, const char * name)
    break;
   }
  }
- mutex_unlock(&dir->d_inode->i_mutex);
+out_unlock:
+ mutex_unlock(&inode->i_mutex);

  if (!found)
   return -ENOENT;

+ sysfs_prune_shadow_sd(parent_sd);
  sysfs_drop_dentry(sd);
  sysfs_deactivate(sd);
```

```
  sysfs_put(sd);
diff --git a/fs/sysfs/symlink.c b/fs/sysfs/symlink.c
index ff605d3..ae2129c 100644
--- a/fs/sysfs/symlink.c
+++ b/fs/sysfs/symlink.c
@@ -15,8 +15,11 @@ static int object_depth(struct sysfs_dirent *sd)
 {
  int depth = 0;

- for (; sd->s_parent; sd = sd->s_parent)
+ for (; sd->s_parent; sd = sd->s_parent) {
+  if (sd->s_type == SYSFS_SHADOW)
+   continue;
   depth++;
+ }

  return depth;
 }
@@ -25,17 +28,24 @@ static int object_path_length(struct sysfs_dirent * sd)
 {
  int length = 1;

- for (; sd->s_parent; sd = sd->s_parent)
+ for (; sd->s_parent; sd = sd->s_parent) {
+  if (sd->s_type == SYSFS_SHADOW)
+   continue;
   length += strlen(sd->s_name) + 1;
+ }

  return length;
 }

 static void fill_object_path(struct sysfs_dirent *sd, char *buffer, int length)
 {
+ int cur;
  --length;
  for (; sd->s_parent; sd = sd->s_parent) {
-  int cur = strlen(sd->s_name);
+  if (sd->s_type == SYSFS_SHADOW)
+   continue;
+
+  cur = strlen(sd->s_name);

   /* back up enough to print this bus id with '/' */
   length -= cur;
@@ -67,7 +77,7 @@ static int sysfs_add_link(struct sysfs_dirent * parent_sd, const char * name,
 int sysfs_create_link(struct kobject * kobj, struct kobject * target, const char * name)
 {
```

```
  struct dentry *dentry = NULL;
- struct sysfs_dirent *parent_sd = NULL;
+ struct sysfs_dirent *parent_sd;
  struct sysfs_dirent *target_sd = NULL;
  int error = -EEXIST;

@@ -81,7 +91,6 @@ int sysfs_create_link(struct kobject * kobj, struct kobject * target, const char

  if (!dentry)
   return -EFAULT;
- parent_sd = dentry->d_fsdata;

  /* target->dentry can go away beneath us but is protected with
   * kobj_sysfs_assoc_lock.  Fetch target_sd from it.
@@ -95,7 +104,10 @@ int sysfs_create_link(struct kobject * kobj, struct kobject * target, const
char
   return -ENOENT;

  mutex_lock(&dentry->d_inode->i_mutex);
- if (!sysfs_dirent_exist(dentry->d_fsdata, name))
+ parent_sd = sysfs_resolve_for_create(target, dentry);
+ if (IS_ERR(parent_sd))
+  error = PTR_ERR(parent_sd);
+ else if (!sysfs_dirent_exist(parent_sd, name))
   error = sysfs_add_link(parent_sd, name, target_sd);
  mutex_unlock(&dentry->d_inode->i_mutex);

@@ -114,7 +126,7 @@ int sysfs_create_link(struct kobject * kobj, struct kobject * target, const
char

 void sysfs_remove_link(struct kobject * kobj, const char * name)
 {
- sysfs_hash_and_remove(kobj->dentry,name);
+ sysfs_hash_and_remove(kobj, kobj->dentry,name);
 }

 static int sysfs_get_target_path(struct sysfs_dirent * parent_sd,
diff --git a/fs/sysfs/sysfs.h b/fs/sysfs/sysfs.h
index 6258462..1c4d20f 100644
--- a/fs/sysfs/sysfs.h
+++ b/fs/sysfs/sysfs.h
@@ -47,6 +47,10 @@ struct sysfs_dirent {
 extern struct vfsmount * sysfs_mount;
 extern struct kmem_cache *sysfs_dir_cachep;

+extern void sysfs_prune_shadow_sd(struct sysfs_dirent *sd);
+struct sysfs_dirent *sysfs_resolve_for_create(struct kobject *, struct dentry *);
+struct sysfs_dirent *sysfs_resolve_for_remove(struct kobject *, struct dentry *);
```

```
+
 extern struct sysfs_dirent *sysfs_get_active(struct sysfs_dirent *sd);
 extern void sysfs_put_active(struct sysfs_dirent *sd);
 extern struct sysfs_dirent *sysfs_get_active_two(struct sysfs_dirent *sd);
@@ -66,7 +70,7 @@ extern void sysfs_attach_dirent(struct sysfs_dirent *sd,
     struct dentry *dentry);

 extern int sysfs_add_file(struct dentry *, const struct attribute *, int);
-extern int sysfs_hash_and_remove(struct dentry * dir, const char * name);
+extern int sysfs_hash_and_remove(struct kobject *, struct dentry *, const char *);
 extern struct sysfs_dirent *sysfs_find(struct sysfs_dirent *dir, const char * name);

 extern int sysfs_create_subdir(struct kobject *, const char *, struct dentry **);
diff --git a/include/linux/sysfs.h b/include/linux/sysfs.h
index a3fa5b9..3db9b16 100644
--- a/include/linux/sysfs.h
+++ b/include/linux/sysfs.h
@@ -72,11 +72,17 @@ struct sysfs_ops {
  ssize_t (*store)(struct kobject *,struct attribute *,const char *, size_t);
 };

+struct shadow_dir_operations {
+ const void *(*current_tag)(void);
+ const void *(*kobject_tag)(struct kobject *kobj);
+};
+
 #define SYSFS_ROOT  0x0001
 #define SYSFS_DIR  0x0002
 #define SYSFS_KOBJ_ATTR  0x0004
 #define SYSFS_KOBJ_BIN_ATTR 0x0008
 #define SYSFS_KOBJ_LINK  0x0020
+#define SYSFS_SHADOW  0x0040
 #define SYSFS_NOT_PINNED (SYSFS_KOBJ_ATTR | SYSFS_KOBJ_BIN_ATTR |
SYSFS_KOBJ_LINK)
 #define SYSFS_COPY_NAME  (SYSFS_DIR | SYSFS_KOBJ_LINK)

@@ -129,6 +135,8 @@ void sysfs_remove_file_from_group(struct kobject *kobj,

 void sysfs_notify(struct kobject * k, char *dir, char *attr);

+int sysfs_enable_shadowing(struct kobject *, const struct shadow_dir_operations *);
+
 extern int __must_check sysfs_init(void);

 #else /* CONFIG_SYSFS */
@@ -224,6 +232,12 @@ static inline void sysfs_notify(struct kobject * k, char *dir, char *attr)
 {
 }
```

```
+static inline int sysfs_enable_shadowing(struct kobject *kobj,
+    const struct shadow_dir_operations *shadow_ops)
+{
+ return 0;
+}
+
 static inline int __must_check sysfs_init(void)
 {
  return 0;
--
1.5.1.1.181.g2de0
```

_____

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