What's coming up in containers

Cgroup Namespaces

- Virtualize /proc/self/cgroup
- Virtualize new cgroupfs mounts
- (example)

Namespaced file capabilities

* Allow containers to lay down file caps
* Without risking host being tricked

Tag security.nscapability with rootid:

* kuid of ns owner
* kuid of ns root
* either one?
* or just > 1 entries

Allowing container root to write xattr:

* New setfcap system call
* Allow setxattr if kuid matches writer
  * Support through libcap extension
  * Supports tar
  * But cannot detect own kuid if nested
* Kernel just fills in kuid

Over to Seth

Mounts from user namespaces

VFS changes

- s_user_ns
- suid
- File capabilities
- Security modules
- Other miscellaneous changes

FUSE

- Already supports unprivileged mounts
  - But via a suid root helper ...
- User/group id translation
- Process id translation

FUSE: security

- About that suid root helper ...
- nosuid
- allow_other

ext4

- User/group id translation
- Limit privileged mount options
- Journal device access
- Backing store attacks

loopfs

- Psuedo filesystem
- Allows user namespaces to allocate loop devices
- Mount at /dev/loop
- Manage via /dev/loop/loop-control
- Kernel creates device nodes
- Prevent backing file modifications

Questions?