Subject: [PATCH 0/20] Pid namespaces
Posted by Pavel Emelianov on Fri, 10 Aug 2007 11:45:08 GMT
View Forum Message <> Reply to Message

A pid namespace is a "view" of a particular set of tasks on the system. They work in a similar way to filesystem namespaces. A file (or a process) can be accessed in multiple namespaces, but it may have a different name in each. In a filesystem, this name might be /etc/passwd in one namespace, but /chroot/etc/passwd in another.

For processes, a process may have pid 1234 in one namespace, but be pid 1 in another. This allows new pid namespaces to have basically arbitrary pids, and not have to worry about what pids exist in other namespaces. This is essential for checkpoint/restart where a restarted process's pid might collide with an existing process on the system's pid.

In this particular implementation, pid namespaces have a parent-child relationship, just like processes. A process in a pid namespace may see all of the processes in the same namespace, as well as all of the processes in all of the namespaces which are children of its namespace. Processes may not, however, see others which are in their parent's namespace, but not in their own. The same goes for sibling namespaces.

The know issue to be solved in the nearest future is signal handling in the namespace boundary. That is, currently the namespace's init is treated like an ordinary task that can be killed from within an namespace. Ideally, the signal handling by the namespace's init should have two sides: when signaling the init from its namespace, the init should look like a real init task, i.e. receive only those signals, that is explicitly wants to; when signaling the init from one of the parent namespaces, init should look like an ordinary task, i.e. receive any signal, only taking the general permissions into account.

The pid namespace was developed by Pavel Emlyanov and Sukadev Bhattiprolu and we eventually came to almost the same implementation, which differed in some details. This set is based on Pavel's patches, but it includes comments and patches that from Sukadev.

Many thanks to Oleg, who reviewed the patches, pointed out many BUGs and made valuable advises on how to make this set cleaner.

Signed-off-by: Pavel Emelyanov <xemul@openvz.org>
Signed-off-by: Sukadev Bhattiprolu <sukadev@us.ibm.com>
Cc: Oleg Nesterov <oleg@tv-sign.ru>