Subject: Re: [PATCH 0/9] namespaces: Introduction
Posted by Dave Hansen on Fri, 19 May 2006 20:04:49 GMT
View Forum Message <> Reply to Message

On Fri, 2006-05-19 at 08:13 -0700, Andrew Morton wrote:
> snapshot/restart/migration worry me. If they require complete
> serialisation of complex kernel data structures then we have a problem,
> because it means that any time anyone changes such a structure they need to
> update (and test) the serialisation.

The idea of actually serializing kernel data structures keeps me up at night. This is especially true when we already have some method of exporting these structures to userspace.

Take VMAs, for example. Should we have a set of interfaces for saving and restoring VMAs, or should we just make any program which is doing checkpoint/restart use /proc/<pid>/maps on checkpoint and mmap() on restore?

It, of course, isn't that simple. Any interface focused on VMAs inside the kernel will have the serialization issues you describe. I think this is such an approach:

http://git.openvz.org/?p=linux-2.6-openvz;a=blob;f=kernel/cpt/cpt_mm.c
http://git.openvz.org/?p=linux-2.6-openvz;a=blob;f=kernel/cpt/rst_mm.c

However, the proc-maps/mmap approach would require new interfaces to be implemented. There are plenty of attributes not currently readily visible to userspace like VM_NONLINEAR, or resources which are normally inaccessible to userspace like deleted files. Those would need extended user/kernel interfaces.

I know of at least one in-kernel commercial checkpoint/restart product which was relatively well tested with "a certain large DB that uses remap_file_pages()" that never even noticed that it completely missed VM_NONLINEAR support until vm-savvy people saw the code. Scary.

-- Dave