Subject: Re: [PATCH 01/11] SYSCTL: export root and set handling routines
Posted by ebiederm on Mon, 19 Dec 2011 10:14:30 GMT

View Forum Message <> Reply to Message

Stanislav Kinsbursky <skinsbursky@parallels.com> writes:

>> Stanislav Kinsbursky<skinsbursky@parallels.com> writes:
>>
>>> These routines are required for making SUNRPC sysctl's per network namespace
>>> context.
>>
>> Why does sunrpc require it's own sysctl root? You should be able to use
>> the generic per network namespace root and call it good.
>>
>> What makes register_net_sysctl_table and register_net_sysctl_ro_table
>> unsuitable for sunrpc. I skimmed through your patches and I haven't
>> seen anything obvious.
>>
>> Eric
>>
>
> Hello, Eric. Sorry for the lack of information.
> I was considering two ways how to make these sysctl per net ns:
>
> 1) Use register_net_sysctl_table and register_net_sysctl_ro_table as you
> mentioned. This was easy and cheap, but also means, than all user-space
> programs, tuning SUNRPC will be broken (since all sysctl currently located
> in"/proc/sys/sunprc/").

Nope. That is a misunderstanding. register_net_sysctl_table works for
anything under /proc/sys.

> 2) Export sysctl root creation routines and make per-net SUNRPC sysctl
> root. This approach allows to make any part of sysctl tree per namespace context
> and thus leave user-space stuff unchanged.
>
> BTW, NFS and LockD also have it's sysctls ("/proc/sys/fs/nfs/").
> And also because of them I've decided, that it would be better to export SYSCTL
> root creation routines instead of breaking compatibility for all NFS layers by
> moving all sysctl under /proc/sys/net/ directory.
>
> Do you feel that it was a bad decision?

I think it was a misinformed decision.

I fully support not breaking userspace by moving where the sysctls files
are. If something sounds like I am suggesting moving sysctl files there

is a miscommunication somewhere.

The concept of a sysctl root as I had envisioned it and essentially as it is implemented was a per namespace sysctl tree. Those sysctl trees are then unioned together when presented to user space. There should only be one root per namespace.

In practice what this means is that register_net_sysctl_table should work for any sysctl file anywhere under /proc/sys. I think register_net_sysctl_table is the right solution for your problem. The only possible caveat I can think of is you might hit Al's performance optimizations and need to create a common empty directory first with register_sysctl_paths.

....
That said since I am in the process of rewriting things some of this may change a little bit, but hopefully not in ways that immediately effect the users of register_sysctl_table.

Don't use register_net_sysctl_ro_table. I think what the implementors actually wanted was register_net_sysctl_table(&init_net, ...) and didn't know it.

Don't put subdirectories in your sysctl tables. Use a ctl_path to specify the entire directory where the files should show up. Generally the code is easier to read in that form, and the code is simpler to deal with if we don't have to worry about directories.

Don't play with the sysctl roots. It is my intention to completely kill them off and replace them by moving the per net sysctl tree under /proc/<pid>/sys/. Leaving behind symlinks in /proc/sys/net and I guess ultimately in /proc/sys/sunrpc/ and /proc/sys/fs/nfs... Which actually seems to better describe your mental model.

Thank you for mentioning /proc/sys/fs/nfs. That is a case I hadn't thought about. In thinking about it I see some deficiencies in my rewrite that I need to correct before I push that code.

Eric