Subject: Re: [PATCH 0/16] Pid namespaces
Posted by Herbert Poetzl on Mon, 09 Jul 2007 19:52:41 GMT
View Forum Message <> Reply to Message

On Mon, Jul 09, 2007 at 05:16:17PM +0400, Pavel Emelianov wrote:
> Herbert Poetzl wrote:
> > On Fri, Jul 06, 2007 at 12:01:59PM +0400, Pavel Emelianov wrote:
> >> This is "submition for inclusion" of hierarchical, not kconfig
> >> configurable, zero overheaded ;) pid namespaces.
> >>
> >> The overall idea is the following:
> >>
> >> The namespace are organized as a tree - once a task is cloned
> >> with CLONE_NEWPIDS (yes, I've also switched to it :) the new
> >> namespace becomes the parent's child and tasks living in the
> >> parent namespace see the tasks from the new one. The numerical
> >> ids are used on the kernel-user boundary, i.e. when we export
> >> pid to user we show the id, that should be used to address the
> >> task in question from the namespace we're exporting this id to.
> >
> > how does that behave when:
> >
> > a) the parent dies and gets reaped?
>
> The children are re-parented to the namespace's init.
> Surprised?
>
> > b) the 'spawned' init dies, but other tasks
> >   inside the pid space are still active?
>
> The init's init becomes the namespace's init.

so an init from the parent process is chosen here?
or 'the init' process? or what am I missing here?

> > c) what visibility rules do apply for the
> >   various spaces (including the default host space)?
>
> Each task sees tasks from its namespace and all its children
> namespaces. Yes, each task can see itself as well ;)
>
> >> The main difference from Suka's patches are the following:
> >>
> >> 0. Suka's patches change the kernel/pid.c code too heavy.
> >>   This set keeps the kernel code look like it was without
> >>   the patches. However, this is a minor issue. The major is:
> >>
> >> 1. Suka's approach is to remove the notion of the task's

> >> numerical pid from the kernel at all. The numbers are
> >> used on the kernel-user boundary or within the kernel but
> >> with the namespace this nr belongs to. This results in
> >> massive changes of struct's members fro int pid to struct
> >> pid *pid, task->pid becomes the virtual id and so on and
> >> so forth.
> >> My approach is to keep the good old logic in the kernel.
> >> The task->pid is a global and unique pid, find_pid() finds
> >> the pid by its global id and so on. The virtual ids appear
> >> on the user-kernel boundary only. Thus drivers and other
> >> kernel code may still be unaware of pids unless they do not
> >> communicate with the userspace and get/put numerical pids.
> >
> > interesting ... not sure that is what kernel folks
> > have in mind though (IIRC, the struct pid change was
> > considered a kernel side cleanup)
>
> That's why I'm sending the patches - to make "kernel folks" make
> a decision. Will we see some patches from VServer team?

unlikely, as we do not require any pid virtualization
except for the init pid (and blend through init)

but I'm worried about the fact that pid spaces will
show up in the host context, which is usually not
what the administrator likes to see ...
(besides the fact that there probably is no way to
tell what processes are real host processes at first
glance, at least not with proper updates to ps and
friends, which might be an option)

> >> And some more minor differences:
> >>
> >> 2. Suka's patches have the limit of pid namespace nesting.
> >> My patches do not.
> >>
> >> 3. Suka assumes that pid namespace can live without proc mount
> >> and tries to make the code work with pid_ns->proc_mnt change
> >> from NULL to not-NULL from times to times.
> >> My code calls the kern_mount() at the namespace creation and
> >> thus the pid_namespace always works with proc.
> >
> > shouldn't that be done by userspace instead?
>
> It can be. But when the namespace is being created there's no
> any userspace in it yet.

I'm not talking about the 'userspace inside the space'

I'm talking about the userspace creating the space
(what if I do not want to have any proc mount?)

> >> There are some small issues that I can describe if someone is
> >> interested.
> >>
> >> The tests like nptl perf, unixbench spawn, getpid and others
> >> didn't reveal any performance degradation in init_namespace
> >> with the RHEL5 kernel .config file. I admit, that different
> >> .config-s may show that patches hurt the performance, but the
> >> intention was *not* to make the kernel work worse with popular
> >> distributions.
> >>
> >> This set has some ways to move forward, but this is some kind
> >> of a core, that do not change the init_pid_namespace behavior
> >> (checked with LTP tests) and may require some hacking to do
> >> with the namespaces only.
> >
> > TIA,
> > Herbert
>
> BTW, why did you remove Suka and Serge from Cc?

once again, I do NOT remove anybody unless explicitely
asked to do so, but I can do nothing against a
broken mailing list ...

(so please go figure where the CC got lost, if you
are sure you added it in the first place)

here the headers:

>From containers-bounces@lists.linux-foundation.org Fri Jul 6 10:03:04 2007
Return-Path: containers-bounces@lists.linux-foundation.org
X-Original-To: herbert@13thfloor.at
Delivered-To: herbert@13thfloor.at
Received: from smtp2.linux-foundation.org (smtp2.linux-foundation.org
+[207.189.120.14])
	(using TLSv1 with cipher DHE-RSA-AES256-SHA (256/256 bits))
	(No client certificate requested)
	by mail.13thfloor.at (Postfix) with ESMTP id 18186702C9
	for <herbert@13thfloor.at>; Fri, 6 Jul 2007 10:02:35 +0200 (CEST)
Received: from murdock.linux-foundation.org (localhost [127.0.0.1])
	by smtp2.linux-foundation.org (8.13.5.20060308/8.13.5/Debian-3ubuntu1.1)+with ESMTP id
l6682UJJ009593;
	Fri, 6 Jul 2007 01:02:32 -0700
Received: from relay.sw.ru (mailhub.sw.ru [195.214.233.200])
	by smtp2.linux-foundation.org

(8.13.5.20060308/8.13.5/Debian-3ubuntu1.1) with ESMTP id
l6682PXL009585
(version=TLSv1/SSLv3 cipher=DHE-RSA-AES256-SHA bits=256 verify=NO)
for <containers@lists.osdl.org>; Fri, 6 Jul 2007 01:02:28 -0700
Received: from [192.168.3.76] ([192.168.3.76])
by relay.sw.ru (8.13.4/8.13.4) with ESMTP id l6681xfW003026;
Fri, 6 Jul 2007 12:02:00 +0400 (MSD)
Message-ID: <468DF6F7.1010906@openvz.org>
Date: Fri, 06 Jul 2007 12:01:59 +0400
From: Pavel Emelianov <xemul@openvz.org>
User-Agent: Thunderbird 1.5 (X11/20060317)
MIME-Version: 1.0
To: Andrew Morton <akpm@osdl.org>
Content-Type: text/plain; charset=ISO-8859-1
Content-Transfer-Encoding: 7bit
Received-SPF: pass (localhost is always allowed.)
X-Spam-Status: No, hits=-3.923 required=5
+tests=AWL,BAYES_00,OSDL_HEADER_SUBJECT_BRACKETED
X-Spam-Checker-Version: SpamAssassin 3.1.0-osdl_revision__1.12__
X-MIMEDefang-Filter: osdl$Revision: 1.181 $
X-Scanned-By: MIMEDefang 2.53 on 207.189.120.22
Cc: Kirill Korotaev <dev@openvz.org>,
Linux Kernel Mailing List <linux-kernel@vger.kernel.org>,
"Eric W. Biederman" <ebiederm@xmission.com>,
Linux Containers <containers@lists.osdl.org>
Subject: [PATCH 0/16] Pid namespaces
X-BeenThere: containers@lists.linux-foundation.org
X-Mailman-Version: 2.1.5
Precedence: list
List-Id: Linux Containers <containers.lists.linux-foundation.org>
List-Unsubscribe:
+<https://lists.linux-foundation.org/mailman/listinfo/containers>,
+<mailto:containers-request@lists.linux-foundation.org?subject=unsubscribe>
List-Archive: <http://lists.linux-foundation.org/pipermail/containers>
List-Post: <mailto:containers@lists.linux-foundation.org>
List-Help: <mailto:containers-request@lists.linux-foundation.org?subject=help>
List-Subscribe:
+<https://lists.linux-foundation.org/mailman/listinfo/containers>,
<mailto:containers-request@lists.linux-foundation.org?subject=subscribe>Sender:
containers-bounces@lists.linux-foundation.org
Errors-To: containers-bounces@lists.linux-foundation.org
Status: RO
X-Status: A
Content-Length: 2788
Lines: 64

>
> Pavel

_____

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