

---

Subject: Checkpoint/Restart mini-summit

Posted by [Daniel Lezcano](#) on Tue, 15 Jul 2008 10:49:45 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

Hi all,

Here is a proposition a more detailed agenda for the checkpoint/restart mini-summit. If everybody is ok with it, I will update the wiki.

Comments are welcome :)

Thanks

-- Daniel

=====

The Checkpoint/restart is a very big topic and the time at the mini-summit is short, so I propose a list of document pointers to be read before the mini-summit, so we can address the checkpoint/restart topic directly and save precious time :)

\* Documentation

- \* Zap : [www.ncl.cs.columbia.edu/publications/usenix2007\\_fordist.pdf](http://www.ncl.cs.columbia.edu/publications/usenix2007_fordist.pdf)
- \* Metacluster : [lxc.sourceforge.net/doc/ols2006/lxc-ols2006.pdf](http://lxc.sourceforge.net/doc/ols2006/lxc-ols2006.pdf)
- \* OpenVZ : [http://wiki.openvz.org/Checkpointing\\_and\\_live\\_migration](http://wiki.openvz.org/Checkpointing_and_live_migration)
- \* Checkpoint/Restart technology :  
[http://en.wikipedia.org/wiki/Application\\_checkpointing](http://en.wikipedia.org/wiki/Application_checkpointing)
- \* Virtual Servers and Checkpoint/Restart in Mainstream Linux : Sigops document

-----

This section is about how to prepare the kernel to implement the checkpoint/restart

\* Preparing the kernel internals

- \* Identifying the kernel subsystems
- \* Identifying the process resources
- \* Identifying the frameworks for the CR
- \* Identifying the pieces to target first

Actually, one of the big interrogation is how we transmit the internal state to and from the kernel. There are some little patches doing the checkpoint/restart, taking into account a small part of the kernel resources. Some were made through netlink, others via /proc, others directly with a syscall. There were solutions proposed in the

containers mailing list to use a core dump like file, or a CR filesystem. This section is to discuss about that.

- \* Passing the kernel internal state to/from userspace
  - \* coredump like file
  - \* swap per container
  - \* netlinks
  - \* CR filesystem
  - \* army of different call for the CR (proc, existing syscalls, ...)

The following sections address the checkpoint/restart itself which can be split into three parts: the quiescent point, the checkpoint and the restart.

#### \* Checkpointing / Restarting

- \* Reaching a quiescent point
  - \* for the network
  - \* for the processes
  - \* for the asynchronous IO
- \* Checkpoint
  - \* Preinstalled checkpoint signal handler ?
  - \* syscall ?
  - \* tar of a CR filesystem ?
  - \* monolithic ?
  - \* Dumping processes hierarchy
  - \* Identifying the kernel resource dependencies
  - \* Dumping system wide resources (per namespace ?)
  - \* Dumping process wide resources (from process context ?)  
(Memory is in between system and process resource)
- \* Restarting
  - \* New binary format handler ?
  - \* Identifying the kernel resource dependencies
  - \* Restoring the processes hierarchy
  - \* Restoring system wide resources
  - \* Restoring process wide resources

There is a posix draft, 1003.1m, which specifies a CR semantic. This can be interesting to take it into account and provide an user API based on this specification so we can keep in mind this when we implement the CR in the kernel. I was not able to find the posix draft itself but the man of the CR IRIX implementation sticks to this specification.

- \* Determining the userspace API
  - \* Posix 1003.1m (implementation in IRIX) ?

[http://techpubs.sgi.com/library/tpl/cgi-bin/getdoc.cgi/0650/bks/SGI\\_Admin/CPR\\_OG/sgi\\_html/ch03.html](http://techpubs.sgi.com/library/tpl/cgi-bin/getdoc.cgi/0650/bks/SGI_Admin/CPR_OG/sgi_html/ch03.html)

---

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

---

---

Subject: Re: Checkpoint/Restart mini-summit  
Posted by [ebiederm](#) on Tue, 15 Jul 2008 18:44:40 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

Daniel Lezcano <dlezcano@fr.ibm.com> writes:

> Hi all,  
>  
> Here is a proposition a more detailed agenda for the checkpoint/restart  
> mini-summit. If everybody is ok with it, I will update the wiki.  
>  
> Comments are welcome :)

A reading list is useful, even to help get some ideas circulating before we get there.

Ultimately the technical details will need to be resolve by people discussing things and sending patches back and forth on the mailing lists.

I don't think a detailed agenda is going to get us anywhere. Especially not one focused on the implementation details.

I think we need to start by seeing what we can agree on. Certainly we agree that checkpoint/restart needs to be part of the picture. What are the problems that the linux community can solve with checkpoint/restart.

Then we need to talk about what kind of implementation we want to merge into mainline. How do we sell it, and how do we implement it without affecting long term maintainability.

I think the granularity of our operations, and what state we save is important. I don't think how we save it is important unless it affects one of our requirements.

As for the posix draft and the historical Cray & SGI implementations. They were on the wrong track. They did not have namespace support so they could not in general restore their checkpoints.

There are also a lot of things you have failed to touch on, that I'm not going to go into now.

With any luck the mini-summit before OLS will be the start of a conversation that will go on all week, and continue on the mailing lists.

The real question is how do we coordinate our efforts to build a good linux checkpoint/restart implementation.

- > \* Documentation
- > \* Zap : [www.ncl.cs.columbia.edu/publications/userix2007\\_fordist.pdf](http://www.ncl.cs.columbia.edu/publications/userix2007_fordist.pdf)
- > \* Metacluster : [lxc.sourceforge.net/doc/ols2006/lxc-ols2006.pdf](http://lxc.sourceforge.net/doc/ols2006/lxc-ols2006.pdf)
- > \* OpenVZ : [http://wiki.openvz.org/Checkpointing\\_and\\_live\\_migration](http://wiki.openvz.org/Checkpointing_and_live_migration)
- > \* Checkpoint/Restart technology :  
> [http://en.wikipedia.org/wiki/Application\\_checkpointing](http://en.wikipedia.org/wiki/Application_checkpointing)
- > \* Virtual Servers and Checkpoint/Restart in Mainstream Linux : Sigops
- > document

There is also the classic emacs undump.  
The very simple vmadump from bproc.

Eric

---

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

---

---

Subject: Re: Checkpoint/Restart mini-summit  
Posted by [ebiederm](#) on Tue, 15 Jul 2008 18:54:15 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

Daniel Lezcano <[dlezcano@fr.ibm.com](mailto:dlezcano@fr.ibm.com)> writes:

- > \* Documentation
- > \* Zap : [www.ncl.cs.columbia.edu/publications/userix2007\\_fordist.pdf](http://www.ncl.cs.columbia.edu/publications/userix2007_fordist.pdf)
- > \* Metacluster : [lxc.sourceforge.net/doc/ols2006/lxc-ols2006.pdf](http://lxc.sourceforge.net/doc/ols2006/lxc-ols2006.pdf)
- > \* OpenVZ : [http://wiki.openvz.org/Checkpointing\\_and\\_live\\_migration](http://wiki.openvz.org/Checkpointing_and_live_migration)
- > \* Checkpoint/Restart technology :  
> [http://en.wikipedia.org/wiki/Application\\_checkpointing](http://en.wikipedia.org/wiki/Application_checkpointing)
- > \* Virtual Servers and Checkpoint/Restart in Mainstream Linux : Sigops
- > document

And of course remote fork is a fun related concept.

Eric

---

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

---

---

Subject: Re: Checkpoint/Restart mini-summit  
Posted by [serue](#) on Wed, 16 Jul 2008 15:15:30 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

Quoting Eric W. Biederman (ebiederm@xmission.com):  
> Daniel Lezcano <dlezcano@fr.ibm.com> writes:  
>  
> > Hi all,  
> >  
> > Here is a proposition a more detailed agenda for the checkpoint/restart  
> > mini-summit. If everybody is ok with it, I will update the wiki.  
> >  
> > Comments are welcome :)  
>  
> A reading list is useful, even to help get some ideas circulating  
> before we get there.  
>  
> Ultimately the technical details will need to be resolve by  
> people discussing things and sending patches back and forth  
> on the mailing lists.  
>  
> I don't think a detailed agenda is going to get us anywhere.  
> Especially not one focused on the implementation details.

Right, the whole point of Daniel including a 'reading list' was just so that we can avoid wasting time discussing existing implementations. So he wasn't suggesting that we would be discussing those in detail, in fact quite the opposite.

> I think we need to start by seeing what we can agree on. Certainly we  
> agree that checkpoint/restart needs to be part of the picture. What  
> are the problems that the linux community can solve with  
> checkpoint/restart.  
>  
> Then we need to talk about what kind of implementation we want to  
> merge into mainline. How do we sell it, and how do we implement  
> it without affecting long term maintainability.  
>

> I think the granularity of our operations, and what state we  
> save is important. I don't think how we save it is important  
> unless it affects one of our requirements.  
>  
> As for the posix draft and the historical Cray & SGI implementations.  
> They were on the wrong track. They did not have namespace support  
> so they could not in general restore their checkpoints.  
>  
> There are also a lot of things you have failed to touch on, that  
> I'm not going to go into now.  
>  
> With any luck the mini-summit before OLS will be the start of a  
> conversation that will go on all week, and continue on the mailing  
> lists.

Agreed.

This could be tough to pull off, but if we can walk out of there with a short focused list of coding todos with the intent of pumping out patches by the end of OLS, turning OLS into a bit of a hack-fest, that would imo be great.

(But then that's precisely what I like to do at conferences - sit by some wall and pick something completely new to code, while once in awhile getting up to chat or see a talk. Does that make me anti-social?)

> The real question is how do we coordinate our efforts to build a good  
> linux checkpoint/restart implementation.  
>  
> > \* Documentation  
> > \* Zap : [www.ncl.cs.columbia.edu/publications/userix2007\\_fordist.pdf](http://www.ncl.cs.columbia.edu/publications/userix2007_fordist.pdf)  
> > \* Metacluster : [lxc.sourceforge.net/doc/ols2006/lxc-ols2006.pdf](http://lxc.sourceforge.net/doc/ols2006/lxc-ols2006.pdf)  
> > \* OpenVZ : [http://wiki.openvz.org/Checkpointing\\_and\\_live\\_migration](http://wiki.openvz.org/Checkpointing_and_live_migration)  
> > \* Checkpoint/Restart technology :  
> > [http://en.wikipedia.org/wiki/Application\\_checkpointing](http://en.wikipedia.org/wiki/Application_checkpointing)  
> > \* Virtual Servers and Checkpoint/Restart in Mainstream Linux : Sigops  
> > document  
>  
> There is also the classic emacs undump.  
> The very simple vmadump from bproc.  
>  
> Eric  
>  
\_\_\_\_\_  
> Containers mailing list  
> Containers@lists.linux-foundation.org  
> <https://lists.linux-foundation.org/mailman/listinfo/containers>

---

Subject: Re: Checkpoint/Restart mini-summit  
Posted by [serue](#) on Wed, 16 Jul 2008 15:17:07 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

Quoting Eric W. Biederman (ebiederm@xmission.com):  
> Daniel Lezcano <dlezcano@fr.ibm.com> writes:  
>  
> > \* Documentation  
> > \* Zap : [www.ncl.cs.columbia.edu/publications/userix2007\\_fordist.pdf](http://www.ncl.cs.columbia.edu/publications/userix2007_fordist.pdf)  
> > \* Metacluster : [lxc.sourceforge.net/doc/ols2006/lxc-ols2006.pdf](http://lxc.sourceforge.net/doc/ols2006/lxc-ols2006.pdf)  
> > \* OpenVZ : [http://wiki.openvz.org/Checkpointing\\_and\\_live\\_migration](http://wiki.openvz.org/Checkpointing_and_live_migration)  
> > \* Checkpoint/Restart technology :  
> > [http://en.wikipedia.org/wiki/Application\\_checkpointing](http://en.wikipedia.org/wiki/Application_checkpointing)  
> > \* Virtual Servers and Checkpoint/Restart in Mainstream Linux : Sigops  
> > document  
>  
> And of course remote fork is a fun related concept.

So Daniel, can you add the reading list to the wiki (including Eric's suggestions), maybe above the agenda?

thanks,  
-serge

---

Subject: Re: Checkpoint/Restart mini-summit  
Posted by [Daniel Lezcano](#) on Wed, 16 Jul 2008 15:26:17 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

Serge E. Hallyn wrote:  
> Quoting Eric W. Biederman (ebiederm@xmission.com):  
>> Daniel Lezcano <dlezcano@fr.ibm.com> writes:  
>>  
>>> \* Documentation  
>>> \* Zap : [www.ncl.cs.columbia.edu/publications/userix2007\\_fordist.pdf](http://www.ncl.cs.columbia.edu/publications/userix2007_fordist.pdf)  
>>> \* Metacluster : [lxc.sourceforge.net/doc/ols2006/lxc-ols2006.pdf](http://lxc.sourceforge.net/doc/ols2006/lxc-ols2006.pdf)  
>>> \* OpenVZ : [http://wiki.openvz.org/Checkpointing\\_and\\_live\\_migration](http://wiki.openvz.org/Checkpointing_and_live_migration)  
>>> \* Checkpoint/Restart technology :

>>> [http://en.wikipedia.org/wiki/Application\\_checkpointing](http://en.wikipedia.org/wiki/Application_checkpointing)  
>>> \* Virtual Servers and Checkpoint/Restart in Mainstream Linux : Sigops  
>>> document  
>> And of course remote fork is a fun related concept.  
>  
> So Daniel, can you add the reading list to the wiki (including Eric's  
> suggestions), maybe above the agenda?

Sure.

---

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

---

---

Subject: Re: Checkpoint/Restart mini-summit  
Posted by [ebiederm](#) on Wed, 16 Jul 2008 19:04:48 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

"Serge E. Hallyn" <[serue@us.ibm.com](mailto:serue@us.ibm.com)> writes:

> Right, the whole point of Daniel including a 'reading list' was just so  
> that we can avoid wasting time discussing existing implementations. So  
> he wasn't suggesting that we would be discussing those in detail, in  
> fact quite the opposite.

Right however he was suggesting we discuss how to implement it without  
agreeing on basic principles. Maybe I'm wrong and all of the different  
development groups have a common idea of what needs to be done but  
I would be surprised if that were the case.

>> With any luck the mini-summit before OLS will be the start of a  
>> conversation that will go on all week, and continue on the mailing  
>> lists.  
>  
> Agreed.  
>  
> This could be tough to pull off, but if we can walk out of there with a  
> short focused list of coding todos with the intent of pumping out  
> patches by the end of OLS, turning OLS into a bit of a hack-fest, that  
> would imo be great.

I totally agree.

> (But then that's precisely what I like to do at conferences - sit by  
> some wall and pick something completely new to code, while once in  
> awhile getting up to chat or see a talk. Does that make me  
> anti-social?)



It means you like to code!

Eric

---

Containers mailing list

Containers@lists.linux-foundation.org

<https://lists.linux-foundation.org/mailman/listinfo/containers>

---

---

Subject: Re: Checkpoint/Restart mini-summit

Posted by [serue](#) on Wed, 16 Jul 2008 19:35:38 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

Quoting Eric W. Biederman (ebiederm@xmission.com):

> "Serge E. Hallyn" <serue@us.ibm.com> writes:

>

> > Right, the whole point of Daniel including a 'reading list' was just so  
> > that we can avoid wasting time discussing existing implementations. So  
> > he wasn't suggesting that we would be discussing those in detail, in  
> > fact quite the opposite.

>

> Right however he was suggesting we discuss how to implement it without  
> agreeing on basic principles. Maybe I'm wrong and all of the different  
> development groups have a common idea of what needs to be done but  
> I would be surprised if that were the case.

IMO recent threads have clearly proven that you're right about basic principles. Question is what is the right level to start at first, and how do we go about reaching consensus? I.e. is the first question whether we should do a fully in-kernel checkpoint and restart vs entirely userspace vs a mix, or is there another place we should start?

It sounds like you have some good ideas in any case on where to start so I'm glad you'll be there :)

thanks,  
-serge

> >> With any luck the mini-summit before OLS will be the start of a  
> >> conversation that will go on all week, and continue on the mailing  
> >> lists.

> >

> > Agreed.

> >

> > This could be tough to pull off, but if we can walk out of there with a  
> > short focused list of coding todos with the intent of pumping out  
> > patches by the end of OLS, turning OLS into a bit of a hack-fest, that

> > would imo be great.  
>  
> I totally agree.  
>  
> > (But then that's precisely what I like to do at conferences - sit by  
> > some wall and pick something completely new to code, while once in  
> > awhile getting up to chat or see a talk. Does that make me  
> > anti-social?)  
>  
> It means you like to code!  
>  
> Eric

---

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

---

---

Subject: Re: Checkpoint/Restart mini-summit  
Posted by [ebiederm](#) on Wed, 16 Jul 2008 21:27:59 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

"Serge E. Hallyn" <serue@us.ibm.com> writes:

> Quoting Eric W. Biederman (ebiederm@xmission.com):  
>> "Serge E. Hallyn" <serue@us.ibm.com> writes:  
>>  
>> > Right, the whole point of Daniel including a 'reading list' was just so  
>> > that we can avoid wasting time discussing existing implementations. So  
>> > he wasn't suggesting that we would be discussing those in detail, in  
>> > fact quite the opposite.  
>>  
>> Right however he was suggesting we discuss how to implement it without  
>> agreeing on basic principles. Maybe I'm wrong and all of the different  
>> development groups have a common idea of what needs to be done but  
>> I would be surprised if that were the case.  
>  
> IMO recent threads have clearly proven that you're right about basic  
> principles. Question is what is the right level to start at first, and  
> how do we go about reaching consensus? I.e. is the first question  
> whether we should do a fully in-kernel checkpoint and restart vs  
> entirely userspace vs a mix, or is there another place we should start?

Where all good kernel features start. With the necessary mechanisms in the kernel and the policy in user space.

If we want to replay all of the user space actions to create the environment at the time of the checkpoint we already have all of the

kernel support we need as the applications go to their current state using current system calls.

I think we want something a bit more efficient.

> It sounds like you have some good ideas in any case on where to start so  
> I'm glad you'll be there :)

Will we have a white board or a large piece paper or something we can draw on and talk about?

Eric

---

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

---

---

Subject: Re: Checkpoint/Restart mini-summit  
Posted by [serue](#) on Thu, 17 Jul 2008 02:27:29 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

Quoting Eric W. Biederman (ebiederm@xmission.com):

> "Serge E. Hallyn" <serue@us.ibm.com> writes:

>

> > Quoting Eric W. Biederman (ebiederm@xmission.com):

> > > "Serge E. Hallyn" <serue@us.ibm.com> writes:

> > >

> > > > Right, the whole point of Daniel including a 'reading list' was just so  
> > > that we can avoid wasting time discussing existing implementations. So  
> > > he wasn't suggesting that we would be discussing those in detail, in  
> > > fact quite the opposite.

> > >

> > > Right however he was suggesting we discuss how to implement it without  
> > > agreeing on basic principles. Maybe I'm wrong and all of the different  
> > > development groups have a common idea of what needs to be done but  
> > > I would be surprised if that were the case.

> >

> > IMO recent threads have clearly proven that you're right about basic  
> > principles. Question is what is the right level to start at first, and  
> > how do we go about reaching consensus? I.e. is the first question  
> > whether we should do a fully in-kernel checkpoint and restart vs  
> > entirely userspace vs a mix, or is there another place we should start?

>

> Where all good kernel features start. With the necessary mechanisms in  
> the kernel and the policy in user space.

>  
> If we want to replay all of the user space actions to create the  
> environment at the time of the checkpoint we already have all of the  
> kernel support we need as the applications go to their current state  
> using current system calls.  
>  
> I think we want something a bit more efficient.  
>  
> > It sounds like you have some good ideas in any case on where to start so  
> > I'm glad you'll be there :)  
>  
>  
> Will we have a white board or a large piece paper or something we can draw  
> on and talk about?  
>  
> Eric

Hi C.,

I'm sorry after all the emails I'm not straight on what we're actually getting. We have a U-shaped room, but do we have a blackboard or whiteboard? Is there a speakerphone? Wireless or at least wired internet?

thanks,  
-serge

---

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

---

---

Subject: Re: Checkpoint/Restart mini-summit  
Posted by [Daniel Lezcano](#) on Thu, 17 Jul 2008 16:15:39 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

Eric W. Biederman wrote:  
> Daniel Lezcano <dlezcano@fr.ibm.com> writes:  
>  
>> Hi all,  
>>  
>> Here is a proposition a more detailed agenda for the checkpoint/restart  
>> mini-summit. If everybody is ok with it, I will update the wiki.  
>>  
>> Comments are welcome :)  
>  
> A reading list is useful, even to help get some ideas circulating  
> before we get there.

>

- > Ultimately the technical details will need to be resolved by
- > people discussing things and sending patches back and forth
- > on the mailing lists.
- >
- > I don't think a detailed agenda is going to get us anywhere.
- > Especially not one focused on the implementation details.
- >
- > I think we need to start by seeing what we can agree on. Certainly we
- > agree that checkpoint/restart needs to be part of the picture. What
- > are the problems that the linux community can solve with
- > checkpoint/restart.
- >
- > Then we need to talk about what kind of implementation we want to
- > merge into mainline. How do we sell it, and how do we implement
- > it without affecting long term maintainability.
- >
- > I think the granularity of our operations, and what state we
- > save is important. I don't think how we save it is important
- > unless it affects one of our requirements.
- >
- > As for the posix draft and the historical Cray & SGI implementations.
- > They were on the wrong track. They did not have namespace support
- > so they could not in general restore their checkpoints.
- >
- > There are also a lot of things you have failed to touch on, that
- > I'm not going to go into now.
- >
- > With any luck the mini-summit before OLS will be the start of a
- > conversation that will go on all week, and continue on the mailing
- > lists.
- >
- > The real question is how do we coordinate our efforts to build a good
- > linux checkpoint/restart implementation.
- >
- >> \* Documentation
- >> \* Zap : [www.ncl.cs.columbia.edu/publications/usenix2007\\_fordist.pdf](http://www.ncl.cs.columbia.edu/publications/usenix2007_fordist.pdf)
- >> \* Metacluster : [lxc.sourceforge.net/doc/ols2006/lxc-ols2006.pdf](http://lxc.sourceforge.net/doc/ols2006/lxc-ols2006.pdf)
- >> \* OpenVZ : [http://wiki.openvz.org/Checkpointing\\_and\\_live\\_migration](http://wiki.openvz.org/Checkpointing_and_live_migration)
- >> \* Checkpoint/Restart technology :
- >> [http://en.wikipedia.org/wiki/Application\\_checkpointing](http://en.wikipedia.org/wiki/Application_checkpointing)
- >> \* Virtual Servers and Checkpoint/Restart in Mainstream Linux : Sigops
- >> document
- >
- > There is also the classic emacs undump.
- > The very simple vmadump from bproc.

Thanks Eric for all your comments. I agree the agenda is a little big, I

will reduce it and I will add the points you raised. I have other points from by Oren I will add too, perhaps that will cover more aspect of the discussion.

-- Daniel

---

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

---