Subject: Re: [PATCH 2/2] dm-ioband: I/O bandwidth controller v0.0.4: Document
Posted by Hirokazu Takahashi on Mon, 28 Apr 2008 06:49:15 GMT
View Forum Message <> Reply to Message

Hi,

> > +What's dm-ioband all about?
> > +
> > + dm-ioband is an I/O bandwidth controller implemented as a device-mapper
> > + driver. Several jobs using the same physical device have to share the
> > + bandwidth of the device. dm-ioband gives bandwidth to each job according
> > + to its weight, which each job can set its own value to.
> > +
> > + At this time, a job is a group of processes with the same pid or pgrp or
> > + uid. There is also a plan to make it support cgroup. A job can also be a
> > + virtual machine such as KVM or Xen.
>
> Most writes are performed by pdflush, kswapd, etc. This will lead to large
> inaccuracy.
>
> It isn't trivial to fix. We'd need deep, long tracking of ownership
> probably all the way up to the pagecache page. The same infrastructure
> would be needed to make Sergey's "BSD acct: disk I/O accounting" vaguely
> accurate. Other proposals need it, but I forget what they are.

I'm working on this topic and I've posted patches once.
http://lwn.net/Articles/273802/

The current linux kernel already has the memory subsystem of cgroup.
You can determine which page is owned by which cgroup with it.
I realized we could easily implement "tracking mechanism of block I/O"
on the feature of this memory subsystem.

When you have an I/O request, it requires data transfer between a certain
page and certain sectors in a disk. This means every I/O requests has
a target page, so you can find out the owner of the page using this
feature. I think it will be okay to assume that the owner of the page
is the owner of the I/O request to the page.

I have a plan on porting the patch to the latest version of linux.
I'm going to make it general, so not only dm-ioband but also OpenVz
team and the NEC team, which are is working on having the CFQ scheduler
control the bandwidth, will be able to make use of this feature.

But it will take some time to port it since the code of the memory
subsystem has been being modified a lot recently.

Thank you,

Hirokazu Takahashi.

_____

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