Subject: RE: [RFC][v2][patch 0/12][CFQ-cgroup]Yet another I/O bandwidth controlling subsystem for CGroups bas
Posted by Satoshi UCHIDA on Fri, 09 May 2008 10:17:44 GMT
View Forum Message <> Reply to Message

Hi, Ryo-San.
Thank you for your test results.

In the test #2 and #3, did you use direct write?
I guess you have used the non-direct write I/O (using cache).

CFQ I/O scheduler was extended in my and Vasily's controllers so that both controllers inherit the features of CFQ.

The current CFQ I/O scheduler cannot control non-direct write I/Os.
This main cause is a cache system.
Bio data is created by special daemon process, such as pdflush or kswapd, for the write I/O using cache.
Therefore, many non-direct write I/Os will belong to one of cgroup (perhaps, root of cgroup).

We consider that this problem should be resolved by fixing cache system.
Specifically, I/Os created by collection of cache pages belong to I/O-context for task which wrote data to cache.
This resolution has a problem.
 * Who is the owner of cache page?
    Cache is reused by many tasks.
    Therefore, it is difficult to decide owner.

In the test #3, It seems that system could control I/Os only among read (sdc2 and sdc3).
Therefore, your test shows that our controller can control I/O without above problem.

Meanwhile, I'm very interested in the result of your test #2.
In the non-direct write I/O, performance will be influenced to task scheduling and sequence of output pages.
Therefore, non-direct write I/O will be fair in the current default task scheduler.
However, your result shows almost fair in Vasilly's controller, whereas non fair in ours.
I'm just wondering if this is an accidental result or an usual result.

Thanks,
  Satoshi UCHIDA.

> -----Original Message-----
> From: Ryo Tsuruta [mailto:ryov@valinux.co.jp]
> Sent: Friday, April 25, 2008 6:55 PM
> To: s-uchida@ap.jp.nec.com; vtaras@openvz.org

> Cc: linux-kernel@vger.kernel.org;
> containers@lists.linux-foundation.org; axboe@kernel.dk;
> tom-sugawara@ap.jp.nec.com; m-takahashi@ex.jp.nec.com; devel@openvz.org
> Subject: Re: [RFC][v2][patch 0/12][CFQ-cgroup]Yet another I/O bandwidth
> controlling subsystem for CGroups based on CFQ
>
> Hi,
>
> I report benchmark results of the following I/O bandwidth controllers.
>
> From: Vasily Tarasov <vtaras@openvz.org>
> Subject: [RFC][PATCH 0/9] cgroups: block: cfq: I/O bandwidth
> controlling subsystem for CGroups based on CFQ
> Date: Fri, 15 Feb 2008 01:53:34 -0500
>
> From: "Satoshi UCHIDA" <s-uchida@ap.jp.nec.com>
> Subject: [RFC][v2][patch 0/12][CFQ-cgroup]Yet another I/O bandwidth
> controlling subsystem for CGroups based on CFQ
> Date: Thu, 3 Apr 2008 16:09:12 +0900
>
> The test procedure is as follows:
> o Prepare 3 partitions sdc2, sdc3 and sdc4.
> o Run 100 processes issuing random direct I/O with 4KB data on each
> partitions.
> o Run 3 tests:
> #1 issuing read I/O only.
> #2 issuing write I/O only.
> #3 sdc2 and sdc3 are read, sdc4 is write.
> o Count up the number of I/Os which have done in 60 seconds.
>
> Unfortunately, both bandwidth controllers didn't work as I expected,
> On the test #3, the write I/O ate up the bandwidth regardless of the
> specified priority level.
>
> Vasily's scheduler
> The number of I/Os (percentage to total I/Os)
>
> ---
> | partition | sdc2 | sdc3 | sdc4 | total
> |
> | priority | 7(highest) | 4 | 0(lowest) | I/Os
> |
>
> |---+---+---+---|---
> |
> | #1 read | 3620(35.6%) | 3474(34.2%) | 3065(30.2%) | 10159
> |
> | #2 write | 21985(36.6%) | 19274(32.1%) | 18856(31.4%) | 60115

> |
> | #3 read&write | 5571(7.5%) | 3253(4.4%) | 64977(88.0%) | 73801
> |
>
> ---
>
> Satoshi's scheduler
> The number of I/Os (percentage to total I/O)
>
> ---
> | partition | sdc2 | sdc3 | sdc4 | total
> |
> | priority | 0(highest) | 4 | 7(lowest) | I/Os
> |
>
> |---+---+---+---|---
> |
> | #1 read | 4523(47.8%) | 3733(39.5%) | 1204(12.7%) | 9460
> |
> | #2 write | 65202(59.0%) | 35603(32.2%) | 9673(8.8%) | 110478
> |
> | #3 read&write | 5328(23.0%) | 4153(17.9%) | 13694(59.1%) | 23175
> |
>
> ---
>
> I'd like to see other benchmark results if anyone has.
>
> Thanks,
> Ryo Tsuruta