Subject: Re: RSS controller v2 Test results (lmbench)
Posted by Lee Schermerhorn on Mon, 21 May 2007 13:53:34 GMT
View Forum Message <> Reply to Message

On Fri, 2007-05-18 at 09:37 +0530, Balbir Singh wrote:
> Rik van Riel wrote:
> > Balbir Singh wrote:
> >
> >> A meaningful container size does not hamper performance. I am in the
> >> process
> >> of getting more results (with varying container sizes). Please let me
> >> know
> >> what you think of the results? Would you like to see different
> >> benchmarks/
> >> tests/configuration results?
> >
> > AIM7 results might be interesting, especially when run to crossover.
> >
>
> I'll try and get hold of AIM7, I have some AIM9 results (please
> see the attachment, since the results overflow 80 columns, I've
> attached them).
>
> > OTOH, AIM7 can make the current VM explode spectacularly :)
> >
> > I saw it swap out 1.4GB of memory in one run, on my 2GB memory test
> > system. That's right, it swapped out almost 75% of memory.
> >
>
> This would make a good test case for the RSS and the unmapped page
> cache controller. Thanks for bringing it to my attention.
>
> > Presumably all the AIM7 processes got stuck in the pageout code
> > simultaneously and all decided they needed to swap some pages out.
> > However, the shell got stuck too so I could not get sysrq output
> > on time.
> >
>
> oops! I wonder if AIM7 creates too many processes and exhausts all
> memory. I've seen a case where during an upgrade of my tetex on my
> laptop, the setup process failed and continued to fork processes
> filling up 4GB of swap.

Jumping in late, I just want to note that in our investigations, when
AIM7 gets into this situation [non-responsive system], it's because all
cpus are in reclaim, spinning on an anon_vma spin lock. AIM7 forks [10s
of] thousands of children from a single parent, resultings in thousands
of vmas on the anon_vma list. shrink_inactive_list() must walk this

list twice [page_referenced() and try_to_unmap()] under spin_lock for each anon page.

[Aside: Just last week, I encountered a similar situation on the i_mmap_lock for page cache pages running a 1200 user Oracle/OLTP run on a largish ia64 system. Left the system spitting out "soft lockup" messages/stack dumps overnight. Still spitting the next day, so I decided to reboot.]

I have a patch that turns the anon_vma lock into a reader/writer lock that alleviates the problem somewhat, but with 10s of thousands of vmas on the lists, system still can't swap enough memory fast enough to recover.

We've run some AIM7 tests with Rik's "split lru list" patch, both with and without the anon_vma reader/writer lock patch. We'll be posting results later this week. Quick summary: with Rik's patch, AIM performance tanks earlier, as the system starts swapping earlier. However, system remains responsive to shell input. More into to follow.

>
> > I am trying out a little VM patch to fix that now, carefully watching
> > vmstat output. Should be fun...
> >
>
> VM debugging is always fun!

For some definition thereof...

Lee