Subject: Re: [RFC][PATCH 2/7] RSS controller core
Posted by Dave Hansen on Tue, 13 Mar 2007 17:05:33 GMT
View Forum Message <> Reply to Message

On Tue, 2007-03-13 at 03:48 -0800, Andrew Morton wrote:
> If we use a physical zone-based containment scheme: fake-numa,
> variable-sized zones, etc then it all becomes moot. You set up a container
> which has 1.5GB of physial memory then toss processes into it. As that
> process set increases in size it will toss out stray pages which shouldn't
> be there, then it will start reclaiming and swapping out its own pages and
> eventually it'll get an oom-killing.

I was just reading through the (comprehensive) thread about this from last week, so forgive me if I missed some of it. The idea is really tempting, precisely because I don't think anyone really wants to have to screw with the reclaim logic.

I'm just brain-dumping here, hoping that somebody has already thought through some of this stuff. It's not a bitch-fest, I promise. :)

How do we determine what is shared, and goes into the shared zones? Once we've allocated a page, it's too late because we already picked. Do we just assume all page cache is shared? Base it on filesystem, mount, ...? Mount seems the most logical to me, that a sysadmin would have to set up a container's fs, anyway, and will likely be doing special things to shared data, anyway (r/o bind mounts :).

There's a conflict between the resize granularity of the zones, and the storage space their lookup consumes. We'd want a container to have a limited ability to fill up memory with stuff like the dcache, so we'd appear to need to put the dentries inside the software zone. But, that gets us to our inability to evict arbitrary dentries. After a while, would containers tend to pin an otherwise empty zone into place? We could resize it, but what is the cost of keeping zones that can be resized down to a small enough size that we don't mind keeping it there? We could merge those "orphaned" zones back into the shared zone. Were there any requirements about physical contiguity? What about minimum zone sizes?

If we really do bind a set of processes strongly to a set of memory on a set of nodes, then those really do become its home NUMA nodes. If the CPUs there get overloaded, running it elsewhere will continue to grab pages from the home. Would this basically keep us from ever being able to move tasks around a NUMA system?

-- Dave

Containers mailing list
Containers@lists.osdl.org
https://lists.osdl.org/mailman/listinfo/containers