Subject: Re: [patch00/05]: Containers(V2)- Introduction
Posted by Paul Jackson on Wed, 20 Sep 2006 20:06:56 GMT
View Forum Message <> Reply to Message

Seth wrote:
> I thought the fake NUMA support still does not work on x86_64 baseline
> kernel. Though Paul and Andrew have patches to make it work.

It works. Having long zonelists where one expects to have to scan a
long way down the list has a performance glitch, in the
get_page_from_freelist() code sucks. We don't want to be doing a
linear scan of a long list on this code path.

The cpuset_zone_allowed() routine happens to be the most obvious canary
in this linear scan loop (google 'canary in the mine shaft' for the
idiom), so shows up the problem first.

We don't have patches yet to fix this (well, we might, I still haven't
digested the last couple days worth of postings.) But we are persuing
Andrew's suggestion to cache the zone that we found memory on last time
around, so as to dramatically reduce the chance we have to rescan the
entire dang zonelist every time through this code.

Initially these zonelists had been designed to handle the various
kinds of dma, main and upper memory on common PC architectures, then
they were (ab)used to handle multiple Non-Uniform Memory Nodes (NUMA)
on bigger boxen. So it is not entirely surprising that we hit a
performance speed bump when further (ab)using them to handle multiple
Uniform sub-nodes as part of a memory containerization effort. Each
different kind of use hits these algorithms and data structures
differently.

It seems pretty clear by now that we will be able to pave over this
speed bump without doing any major reconstruction.

--

I won't rest till it's the best ...
Programmer, Linux Scalability
Paul Jackson <pj@sgi.com> 1.925.600.0401