

Re: Tuning for PostgreSQL Database

Source: <http://unix.derkeiler.com/Mailing-Lists/FreeBSD/performance/2003-07/0080.html>

From: Terry Lambert (tlambert2_at_mindspring.com)

Date: 07/24/03

Date: Wed, 23 Jul 2003 23:55:25 -0700

To: "Jim C. Nasby" <jim@nasby.net>

"Jim C. Nasby" wrote:

[... quote of me and quote of Matt Dillon's "Blue Prints" article ...]

> *The question I have is: can pages in the inactive queue be used as disk
> cache?*

The answer is "yes, they can be reactivated and written to before they are flushed if soft updates is enabled" and "yes, they can be reactivated and read (but not written) to before they are flushed if soft updates is not enabled".

In general, this only happens for data pages, which is to say, the pages containing user file data. Pages containing FS metadata are specifically considered as "write through" or "virtually write through".

It doesn't happen for data pages, if they are explicitly fsync'ed to ensure write ordering is guaranteed.

Metadata pages will be marked as "busy" by the system until they are written out in dependency order, once a write is started on the page in question. Effectively, they are "read-only", and reads do not stall, but new writes stall, until the write completes. This only happens *after* the write hits the block I/O subsystem.

In reality, the pages are treated as copy-on-write, with a blocking semantic to ensure metadata serialization (e.g. if there was a bwrite in progress and a bdwrite was requested, it could go through, but another bdwrite would be blocked until the first finished.

IF there are multiple operations in progress in the same page, AND there are no dependencies between the operations, AND soft updates is enabled, AND the write has been paced on the soft updates clock wheel to be written AND the wheel has not progressed to the point where the write has actually been taken off the wheel and scheduled

freebsd-performance: Re: Tuning for PostgreSQL Database

in the I/O subsystem, THEN the write may be scheduled to occur simultaneously, IF there are no intermediate dependent writes that need to take place.

In other words, if the dependency is "soft", then it can gather any modifications to a single page together, and save I/O operations (or in the case of create-write-delete for a shortlived intermediate file, it can avoid the writes altogether.

All this boils down to one thing: in the normal case, metadata write ordering is implicitly guaranteed in all cases where it is not specifically declined at the time the FS is mounted (via the "async" option, the "noatime" option, etc.), all of which are disabled by default.

> *Or maybe a better question would be: what does each memory category in top mean?*
> *Mem: 365M Active, 1400M Inact, 168M Wired, 76M Cache, 199M Buf, 3008K Free*

Depends on the version of "top" you are running. The statistics we keep are in the "struct vmmeter" in the file /usr/src/sys/sys/vmmeter.h.

The meaning of these statistics varies slightly, over time, so that's not fixed either (but I've seen more changes in "top" than FreeBSD).

The place to look for their meanings is first in the source code for the version of "top" you are running, to see what fields they are using and how/if they are combining them mathematically, and then second, in the code that updates the variables you are interested in (usually meaning code that lives in /usr/src/sys/vm/*.c).

Honestly, if you aren't able to dig the information out, you are not likely to be able to understand the answer the way it was intended to be understood, if someone comes right out and tells you.

Kirk McKusick is rumored to be working on a FreeBSD Internals book, but we are going on 3 years for that rumor. I started one, and I updated it several times in the process, but, frankly, FreeBSD will not stand still long enough for a single person to document it well, and I discontinued work on it at about the 4.6-RELEASE level.

IMO, writing a good book takes at least 2080 hours on the part of the author(s), which is equivalent to a full time job for a year, and it also takes a willingness on the part of technical reviewer(s) to spend a lot of time on the review process, in order for the book to be any good (e.g. I spent probably 200 total hours in the review process on Uresh Vahalia's "UNIX Internals: The New Frontiers" for Prentice Hall's technical editor on that project).

> *Is there anywhere that clearly defines what each queue is, and how it's used?*

freebsd-performance: Re: Tuning for PostGreSQL Database

The source code for a particular version tag does, for a version built from that particular version tag, and probably only that version.

-- Terry

freebsd-performance@freebsd.org mailing list

<http://lists.freebsd.org/mailman/listinfo/freebsd-performance>

To unsubscribe, send any mail to "freebsd-performance-unsubscribe@freebsd.org"