Postgres {on,vs,with} Linux

Andres Freund
PostgreSQL Developer & Committer
Email: andres@anarazel.de
Email: andres.freund@microsoft.com

https://anarazel.de/talks/2026-09-23-kernel-recipes-postgres-linux/postgres-linux.pdf

Past Fun of Postgres on Linux

Past Fun of Postgres on Linux

- "fsyncgate"
- io_uring cross process problems, influencing io_uring features
- Various hangs, crashes, errors
- Investigations around various performance changes of postgres on linux in new releases
- kernel writeback triggers crazy latency
- transparent huge & zone reclaim pages can scale terrible (10+ years ago)

Why me?

- I like OS stuff
- I mostly work on lower level parts of postgres
- Thought skimming all of LKML was a good idea as a ~15 year old
- Happened to be lucky enough to start to know a few folks I can ask
- But: I don't scale, I can't kept up, I'm not that smart

Why aren't more of us interacting?

Why aren't more of us interacting?

- Others don't know linux that well

An overview of Linux for Dummies^WDatabase Developers

at a postgres conference
(e.g. pgconf.dev in Montreal)

Anyone?

Why aren't more of us interacting?

- Others don't know linux that well
- Communication Style
- Contradictory Answers
- Keeping up is **really** hard
  - Can’t even follow one thread.
    - learned about b4 here!
- ?

What's happening in Postgres (only kernel relevant stuff)

What's happening in Postgres

- Slowly moving to threading
  - same memory mapping makes a lot of stuff easier
  - TLB utilization, context switch perf
  - medium to long term goal: Separate connections from processes/threads

What's happening in Postgres

- AIO
  - currently only reads; writes, f[data]sync hopefully in 20)
  - backed by workers or io_uring (others coming)
  - Converting more sources of IO to async
- DIO/O_DIRECT
  - won’t be default

What's happening in Postgres

- Experimenting with atomic writes
  - DIO-only a problem, but work on buffered is happening
  - single atomic write unit limit problematic

What's happening in Postgres

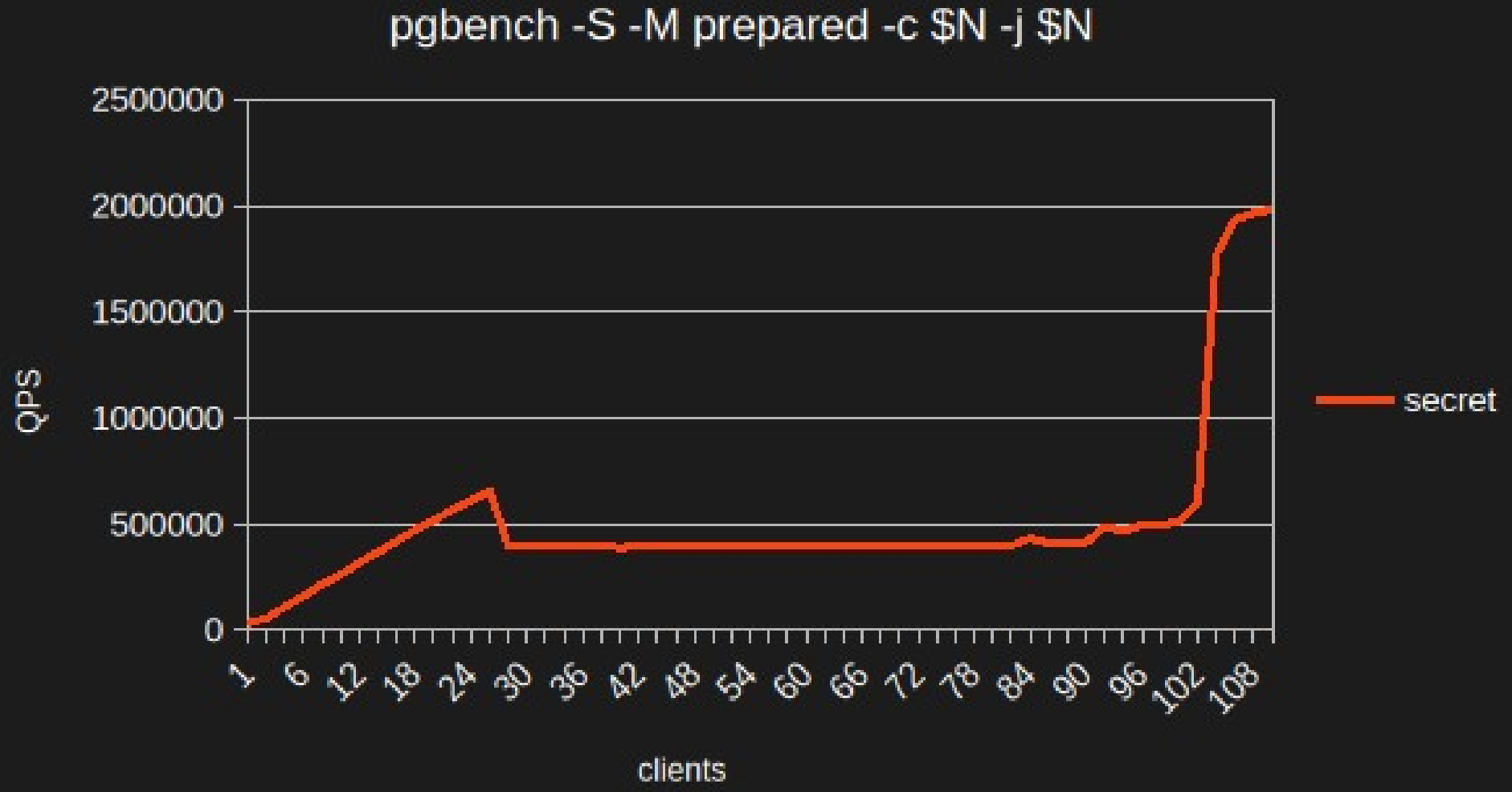
- Experimenting with time slice extensions

Good things happening in Linux!

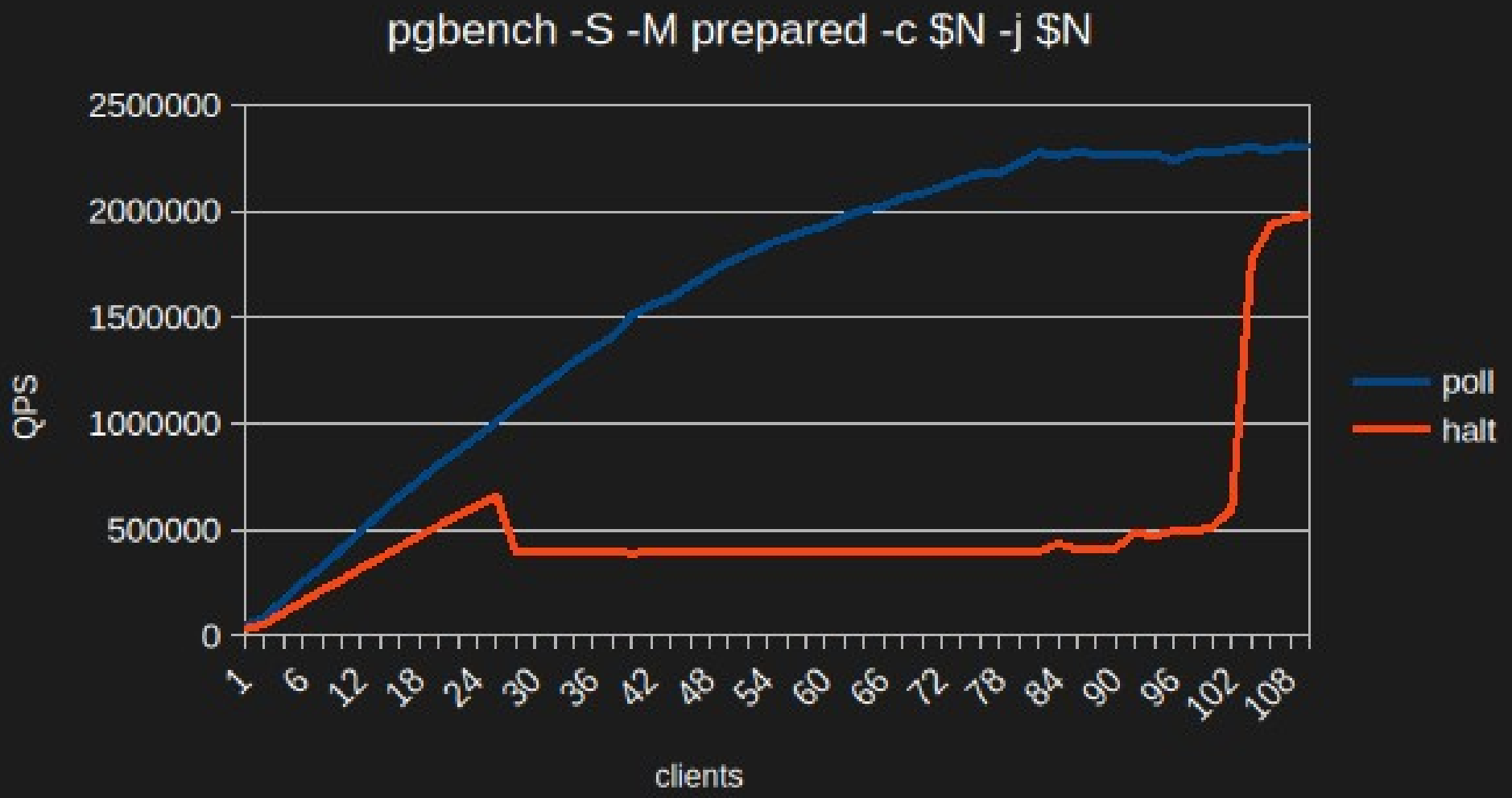
Good things happening in Linux!

- Writeback / reclaim got a *lot* better in the last year or two
  - Like 2-4x faster and much better tail latency
- RWF_ATOMIC is great progress
- io_uring is great
  - Jens has been great at collaborating

What "problems" is postgres seeing



Thanks to Tomas Vondra for the data



Thanks to Tomas Vondra for the data

What "problems" is postgres seeing

- Partially loaded larger machines sometimes perform terrible on context-switch heavy workloads
  - Observed on x86 only so far, wide variety of hardware
  - idle=poll or cpuidle idle-set -D0 makes it often better
  - but those are terrible, preventing real boosting, increase power

What "problems" is postgres seeing

- iTLB slowness
  - 15-20% in some benchmarks
  - READ_ONLY_THP_FOR_FS not enabled on distros & going away
  - upstream progress
  - trailing partially filled % 2MB can’t use huge pages
    - copy in kernel? add hole in file?
  - inconsistent performance results for benchmarking

What "problems" is postgres seeing

- Userspace-controlled directed wakeups would be great
  - Plain futexes don’t wake up lock holder
  - PI Futexes aren’t usable for us and slow
  - Timeslice extension too restrictive
  - Tail latency and oversubscribed systems suffer most
- Futexes
  - 32 byte value quite limiting for us
  - not that fast – the hash lookup shows up heavily in some profiles
  - Directed wakeups – PI futexes

What "problems" is postgres seeing

- Cgroups OOM handling
  - We really like overcommit_memory=2 for prod systems
  - but that doesn’t work in cgroups, unless the memory shortage is global

What "problems" is postgres seeing

- Reserve space on-disk for to-be-checkpointed data
  - fallocate()
  - We use it to reserve space for in-memory-date to avoid ENOSPC
  - often causes fragmentation
  - triggers journal writes (but FALLOC_FL_WRITE_ZEROES will help)
  - doesn’t work on CoW filesystems

What "problems" is postgres seeing

- 1GB huge pages can cause a lot more contention
- io_uring
  - IORING_OP_{READ,WRITE}[V]_FIXED limited to 1GB
  - Complicated heuristics around setting IOSQE_ASYNC
  - Memory locking accounting for rings

Why don't we "task" the kernel devs also employed at the big companies?

How can PG help Linux

- Lots of our problems are more general
- Fewer post-release "BIG REGRESSION, END OF WORLD" "fun"
- Can help you choose better regression benchmarks
- Help another old open source project :)