BOF: Painless kernel - Removing the HZ

Josh Triplett

September 23, 2009

Problem

- "Tickless kernel"? Not really.
- Tickless only when idle
- When running anything, Linux still takes an interrupt HZ times per second

Symptoms

- High tick overhead
  - Measured 8% with HZ=1000, 2% with HZ=250 (YMMV)
- Updating scheduler accounting without rescheduling
- Giant pile of things hanging off `tick_sched_timer`
- Lots of polling
- Slows down real work
- Introduces jitter
- Introduces latency (timer interrupt not preemptible)
- Increases energy consumption

Cheating to see the end result

- Hacked out the "if idle" condition on nohz mode
- Timer interrupt never goes off except when timer explicitly set
- Many known issues: scheduling, CPU accounting, RCU, . . .
- Boots, runs benchmarks

Cheating to see the end result - Before
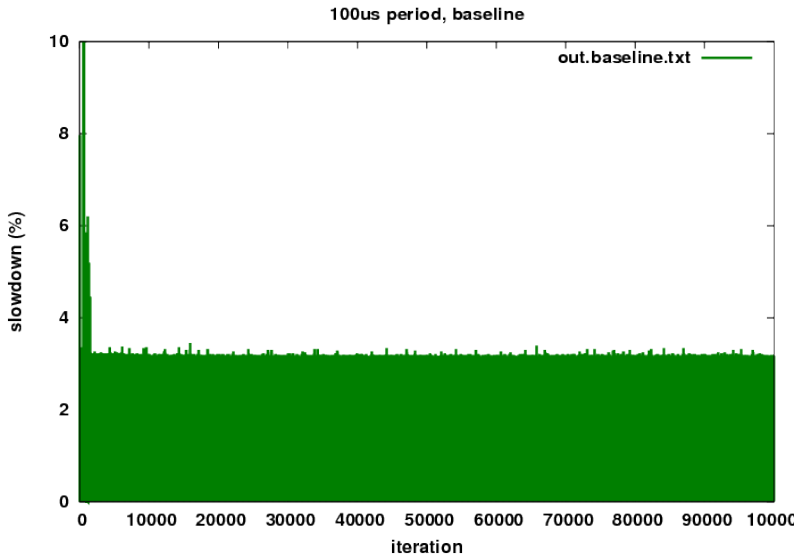
(Graphs by Anton Blanchard)

Cheating to see the end result - After

100us period, user NOHZ
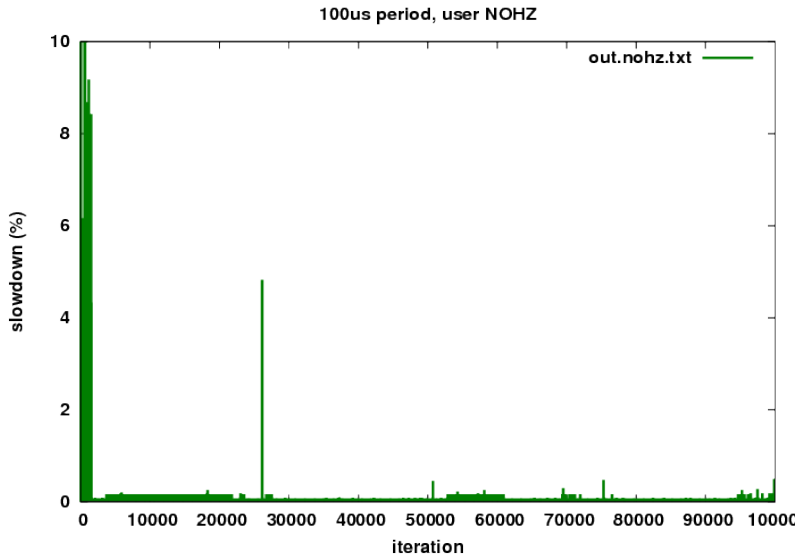
(Graphs by Anton Blanchard)

Real Solutions

- Remove the polling from `tick_sched_timer`
  - Find an appropriate event
  - Set an appropriate one-shot timer
  - Eliminate entirely
- Stop thinking of `jiffies` as free
- Revive the work to make the scheduler itself preempt only when needed