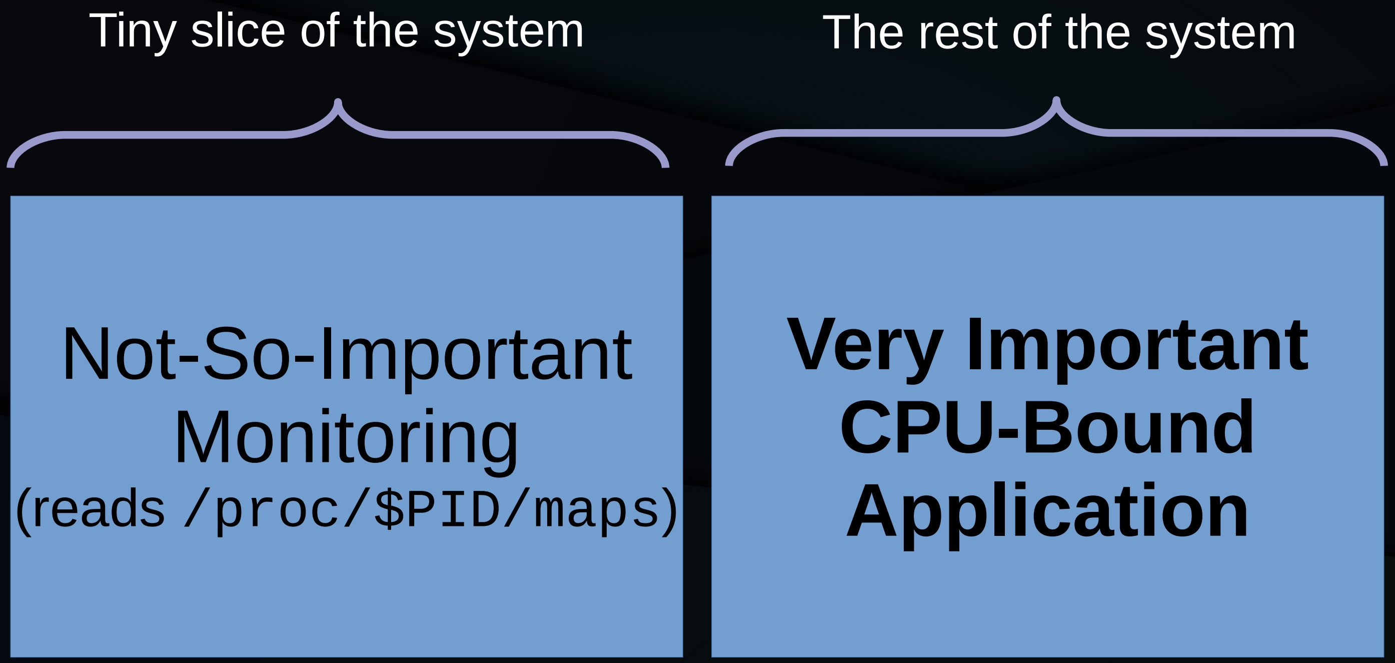
Matthew Wilcox, Oracle & Paul E. McKenney, Facebook

Linux Plumbers Conference Performance & Scalability MC, September 23, 2021

# "cat /proc/$PID/maps": What Could Possibly Go Wrong?

# What Could Possibly Go Wrong???

# What Could Possibly Go Wrong???



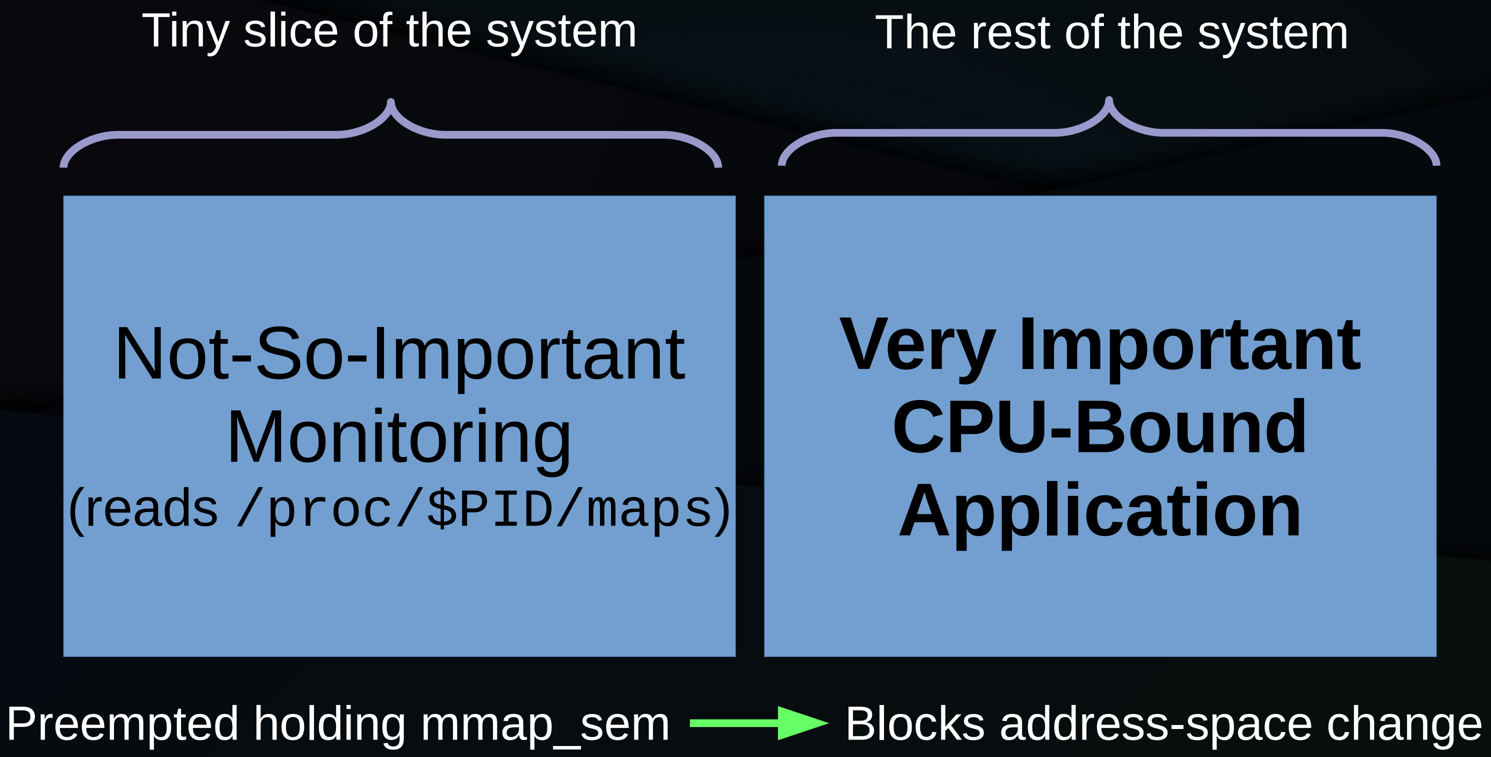
Tiny slice of the system

The rest of the system

Not-So-Important Monitoring (reads `/proc/$PID/maps`)

**Very Important CPU-Bound Application**

# What Could Possibly Go Wrong???



Tiny slice of the system

The rest of the system

Not-So-Important Monitoring (reads `/proc/$PID/maps`)

**Very Important CPU-Bound Application**

Preempted holding mmap_sem → Blocks address-space change

# Exactly How Does This Happen???

1) Not-so-important monitoring (NSIM) acquires mmap_sem to read /proc/$PID/maps

2) Very important CPU-bound application (VICBA) thread A invokes mmap() and blocks write-acquiring mmap_sem

3) VICBA thread B takes a page fault and blocks read-acquiring mmap_sem

4) Other VICBA threads and other unrelated work consume all available CPU, preventing NSIM from running.

5) VICBA threads A & B are blocked indefinitely!!!

# Reproducer

- Problem happens in production, but rarely
- Helpful to have reproducer for testing:
  - One process maps and unmaps a region
  - Another repeatedly scans /proc/$PID/maps
  - Others consume all available CPU

https://github.com/paulmckrcu/proc-mmap_sem-test.git

# 24 Runs of the Reproducer on v5.4

| | Worst-case mmap()/munmap() latency (milliseconds) | | | |
|---:|---:|---:|---:|---:|
| --nbusytasks | Median | Minimum | Maximum | # “hangs” |
| 0 | 0.097 | 0.036 | 0.141 | |
| 1 | 27.296 | 23.932 | 116.081 | |
| 10 | 123.514 | 119.402 | 179.284 | |
| 100 | 357.379 | 307.146 | 1251.496 | |
| 1000 | 8019.600 | 4114.936 | 12020.700 | 23 |

```
./run-proc-vs-map.sh --nsamples 24
```

# VMA Maple Tree

- Tree protected with a spinlock
  - Readers can use RCU
- VMAs are now RCU freed
- Visible inconsistencies are tolerable
  - May see overlapping VMAs
  - May miss newly added VMAs

# Compare With Maple-Tree Prototype

| | Worst-case mmap()/munmap() latency (milliseconds) | | | | | |
|---|---|---|---|---|---|---|
| | V5.4 | | | Maple-Tree Prototype (Jan 2021) | | |
| #Busy | Median | Minimum | Maximum | Median | Minimum | Maximum |
| 0 | 0.039 | 0.036 | 0.088 | 1.329 | 0.991 | 1.825 |
| 1 | 27.037 | 26.955 | 76.058 | 2.007 | 1.742 | 2.017 |
| 2 | 27.577 | 27.243 | 31.574 | 1.797 | 1.571 | 1.870 |

RCU readers decouple /proc/$PID/maps scans from mmap()/munmap()

# Page Table Issue

- /proc/$PID/smaps walks page tables
  - Reports presence of pages
- In RCU mode, can race with unmap
  - And the page tables can be freed under it
- Need to RCU free all page tables