futex2: A new interface

André Almeida
LPC 2020

Current interface

- Last added feature: realtime clock support (2008)
- Minor fixes and internal improvements since then
- Attempted features:
  - Adaptive spinning futexes (2010, 2016)
  - Attached futexes, hash tables per process (2016)
  - Variable sized futexes (2019)
  - Wait on multiple futexes (2019)
  - FUTEX_SWAP (2020)

Current interface

- Legacy features (REQUEUE, BITSET, WAKE_OP, ...)
- “Fragile as hell and horrible to maintain” - Maintainer
- Lack of NUMA awareness
- Determinism: time complexity for RT users

Proposed solution

- Thomas G. and Peter Z.: we should create a new interface
- No more multiplexing: one syscall per operation
- Flags for NUMA, size, shared, clockid
- Code from scratch. Every line reviewed as brand new
- Do one thing and do it well

Proposed interface

- Proposed at mailing list by Peter Z. and Florian W.

```
futex_wait(void *uaddr, unsigned long val,
           unsigned long flags, struct timespec *timo);

futex_wake(void *uaddr, unsigned int nr,
           unsigned long flags);

futex_waitv(struct futex_waitv *waiters[], unsigned int nr_waiters,
            unsigned long flags, struct timespec *timo);

futex_cmp_requeue(void *uaddr1, void *uaddr2, unsigned int nr_wake,
                  unsigned int nr_requeue, u64 cmpval, unsigned long flags);
```

```
struct futex_waitv {
    void *uaddr;
    unsigned long val;
    unsigned long flags;
};
```

Proposed interface

- For a non-NUMA operation `void *uaddr` just points to the integer address
- For a NUMA-aware (FUTEX_NUMA_FLAG) operation, the userspace should use the following struct and set the node_id[-1, MAX_NUM_NODE] in the hint member

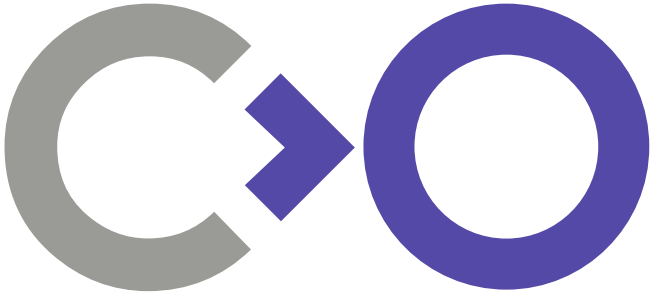
```
struct futex8_numa  {  u8 value;  u8 hint; };
struct futex16_numa { u16 value; u16 hint; };
struct futex32_numa { u32 value; u32 hint; };
struct futex64_numa { u64 value; u64 hint; };
```

- All those value members will be naturally aligned

Development process

- To get all the features, we will need collaboration
- No PI, robust, requeue for now
- Starting with just u32 wait/wake
- What do you think the next steps are?

andre.almeida@collabora.com
LPC chat: andre.almeida.collabora.com



Thank you!