Reduce the Abuse of mmap_sem

Yang Shi

Alibaba

The Problems

- The abuse and misuse of mmap_sem may lead to:
  - Processes are stalled for long time
    - Processes get stuck, hung, …
  - I/O priority inversion
  - Unnecessary contention

Mitigation

- These could be mitigated by:
  - Use trylock when retry acceptable
    - 3b454ad35043 mm: thp: use down_read_trylock() in khugepaged to avoid long block
  - Avoid holding write lock when possible
    - dd2283f2605e mm: mmap: zap pages with read mmap_sem in munmap
  - Avoid holding it when possible
    - https://lwn.net/Articles/754739/ (Speculative page faults)
  - Release the lock earlier when possible
    - https://lwn.net/Articles/766818/ (drop the mmap_sem when doing IO in the fault path)

There is more…

- It is hard to figure out what mmap_sem protects (surprisingly)
  - Rbtree of VMA
    - find_vma()
  - VMA list
    - Lock the whole address space for even touching one byte
  - VMA flags
    - Need hold write lock to update vm_flags
  - Some fields of mm_struct
    - arg_start, arg_end, env_start, env_end, etc, before 4.18

Finer grain lock?

- Will range lock help?
  - Not helpful for the applications which consist of one large VMA
- Per-VMA lock?