# Improvement of PTE page table management

Qi Zheng

ByteDance STE Team

CLK 2021

ByteDance 字节跳动

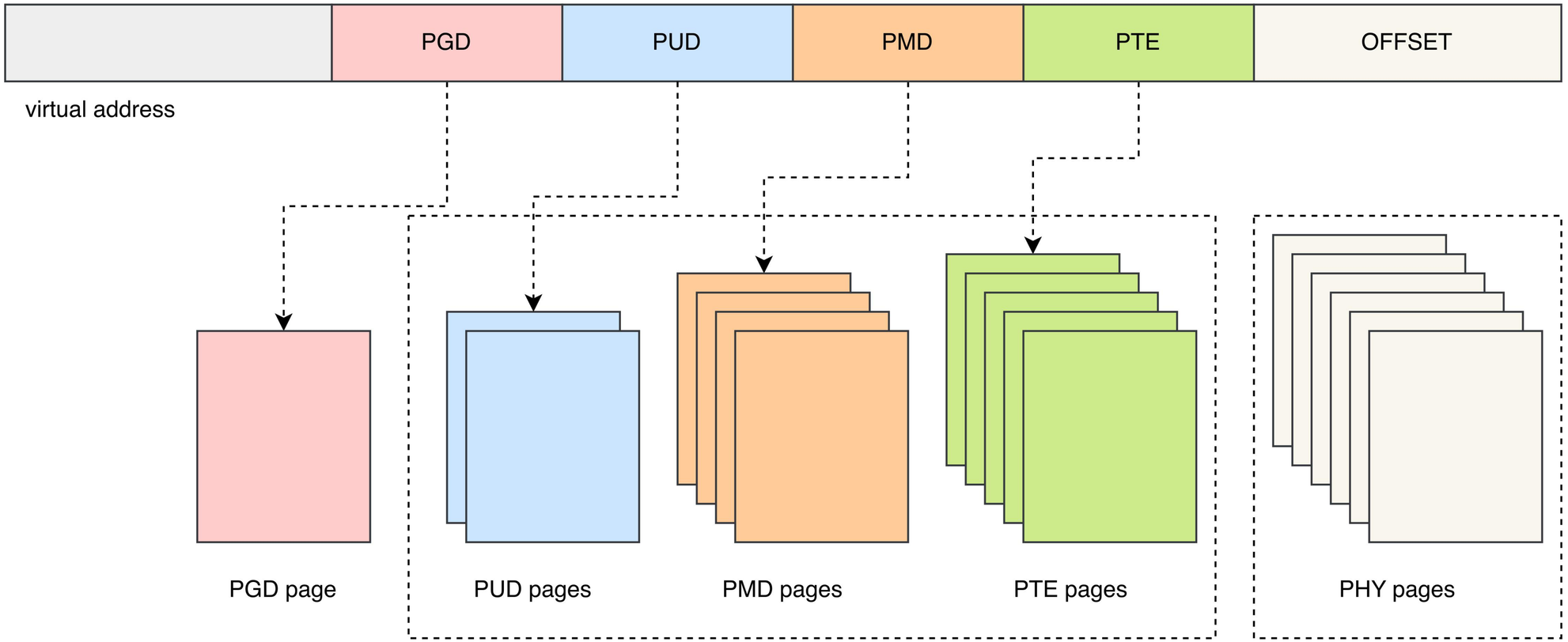
# Agenda

- Background
- Old management
- New management
- Status of upstream work

# Background

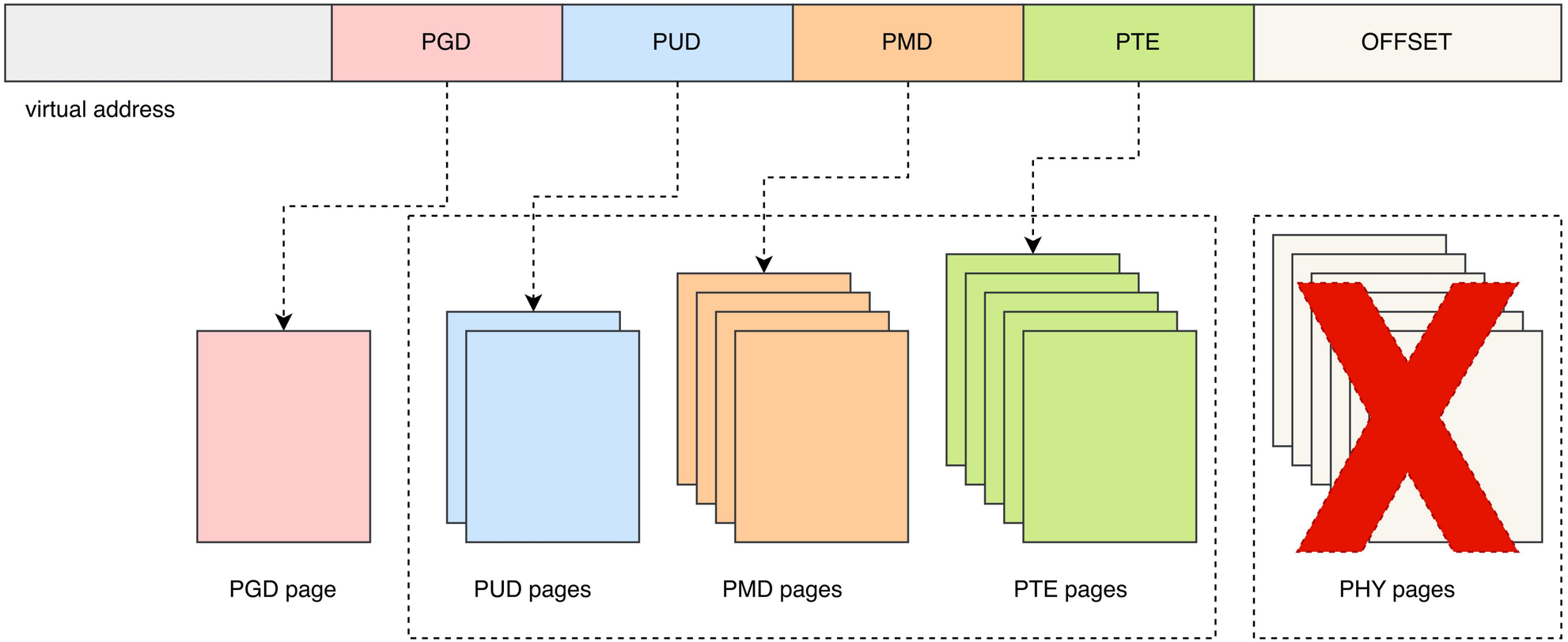
ByteDance字节跳动

# Background




alloc pages after mmap()

# Background




madvise MADV_DONTNEED/MADV_FREE

# Background

Up to **3 MB** **512 GB**

**32 bits** **64 bits**

memory consumption of user PTE pages

# Background

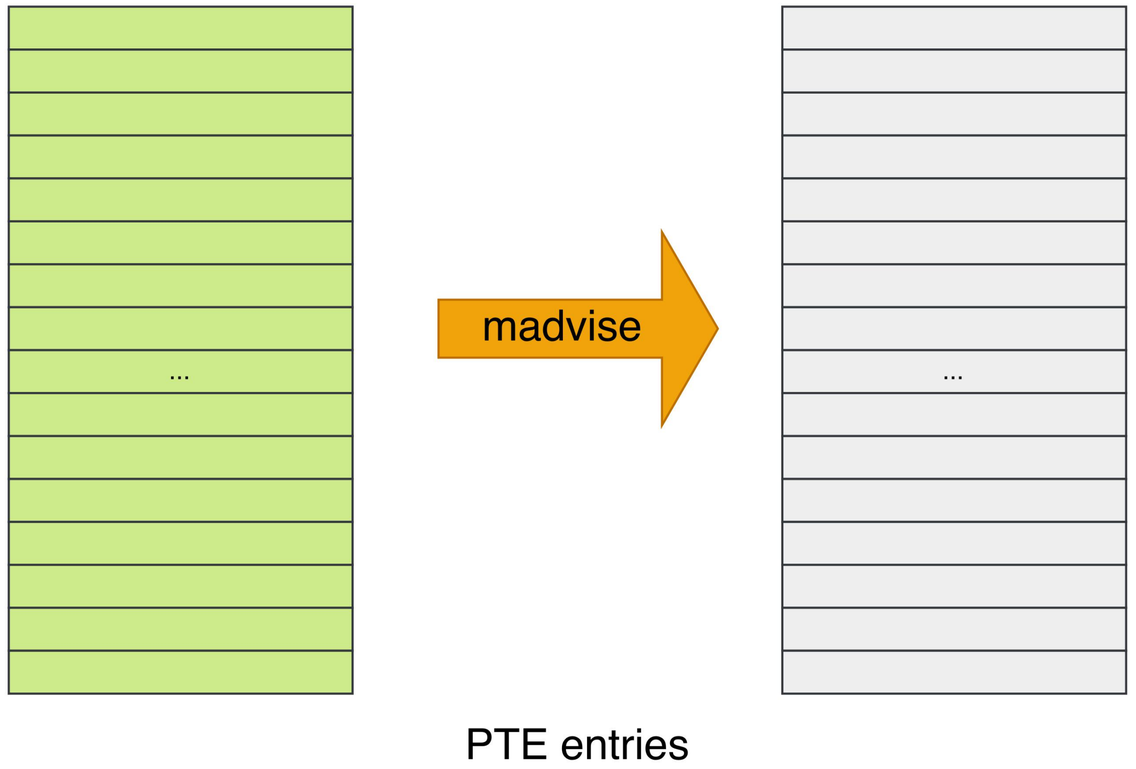
- real case:

VIRT: 55 TB

RES: 590 GB

VmPTE: ~1.2 GB

VmPTE: 110 GB



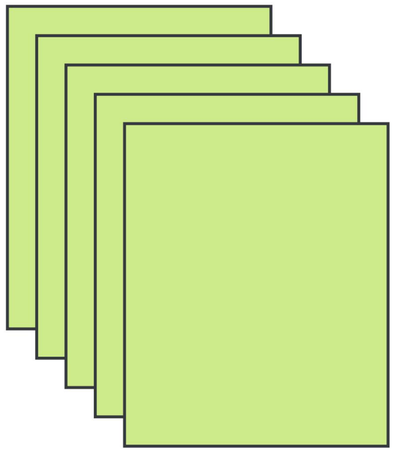

- jemalloc / tcmalloc

# Old management

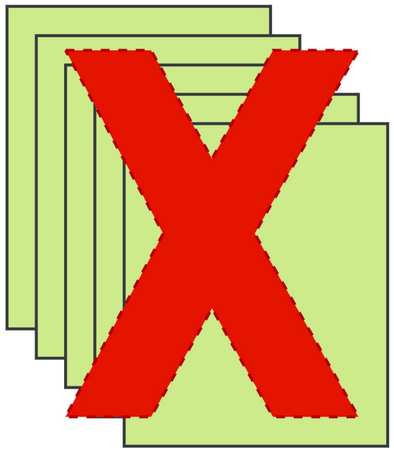
ByteDance字节跳动

# Old management

- event-triggered



page fault



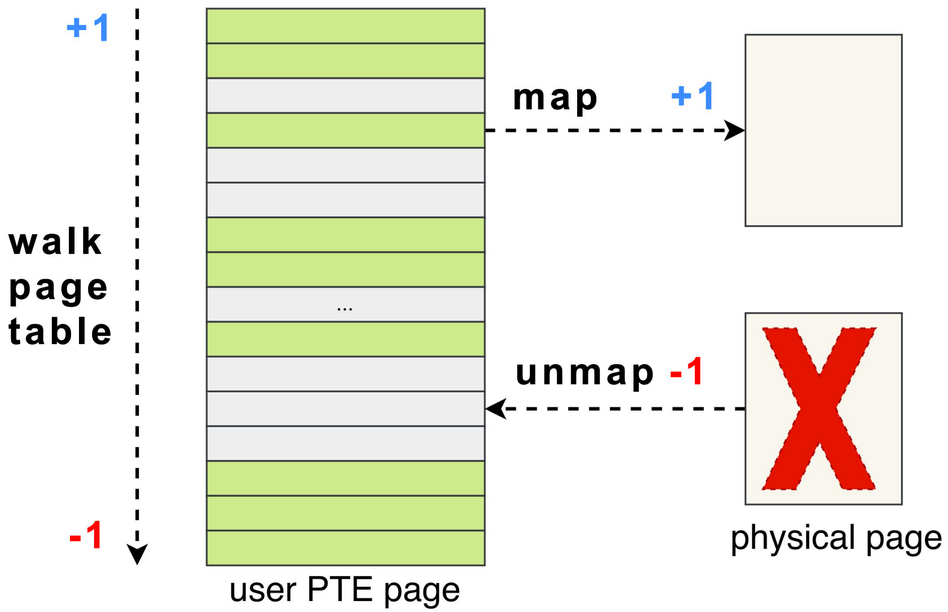
task exit
unmap region

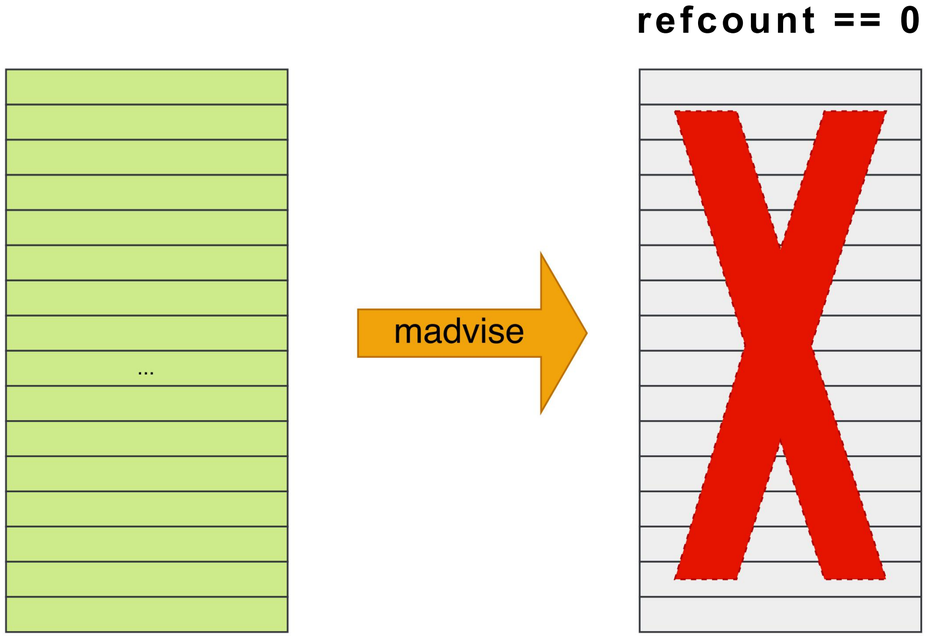
# New management

ByteDance字节跳动

# New management

- introduce refcount for user PTE page table pages
- similar to the refcount of page



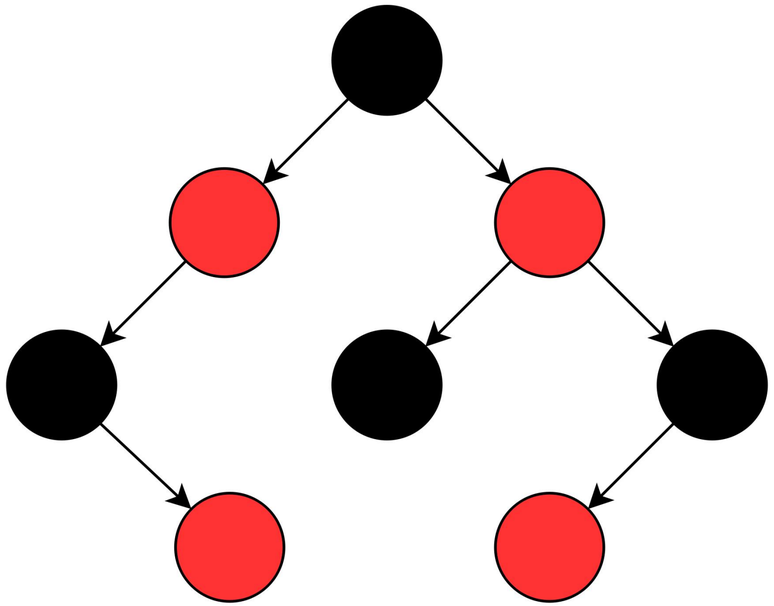
# New management

How to accurately find all the places that need to increase or decrease the reference count?

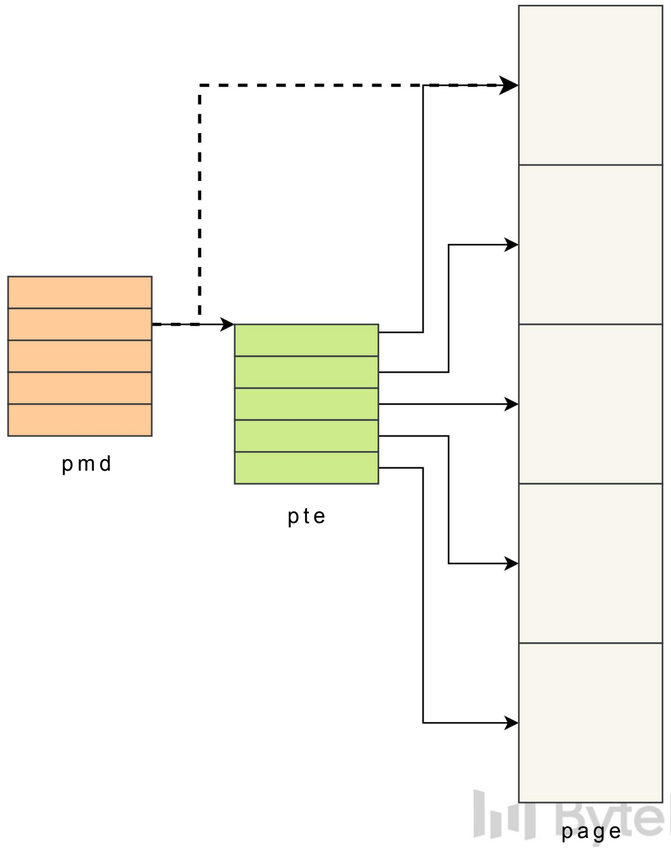
# New management

**before**: as long as you hold the read lock, all things is ok

**now**: even if the read lock is held, the PTE page still can be released



write lock



# New management

- functional testing

| | before | after | |
|---|---|---|---|
| VIRT | 50.0 GB | 50.0 GB | |
| RES | 3.1 MB | 3.6 MB | |
| VmPTE | 102640 kB | 248 kB | saved ~100 MB |

- stability testing

  - LTP: Linux Test Project

# New management

- performance testing

```
root@~# perf stat -e page-faults --repeat 5 ./multi-fault $threads:

threads         before (pf/min)                    after (pf/min)
    1              32,085,255                         31,880,833 (-0.64%)
    8             101,674,967                        100,588,311 (-1.17%)
   16             113,207,000                        112,801,832 (-0.36%)

(The "pfn/min" means how many page faults in one minute.)
```

~1% slower than before

# Status of upstream work

ByteDance字节跳动

# Status of upstream work

- some cleanups and fixes have been merged
  - https://lkml.org/lkml/2021/7/21/135
  - https://lkml.org/lkml/2021/9/23/646
  - https://lkml.org/lkml/2021/9/1/384
- the main patch series is still being updated
  - https://lkml.org/lkml/2021/8/18/1359

# THANKS

ByteDance 字节跳动

kernel trace tools 讨论 2 群

该二维码 7 天内 (10月31日前) 有效，重新进入将更新