Alibaba Cloud
Worldwide Cloud Services Partner

KVM memory cost optimization in Alibaba cloud

Huaitong Han
huaitong.hht@alibaba-inc.com

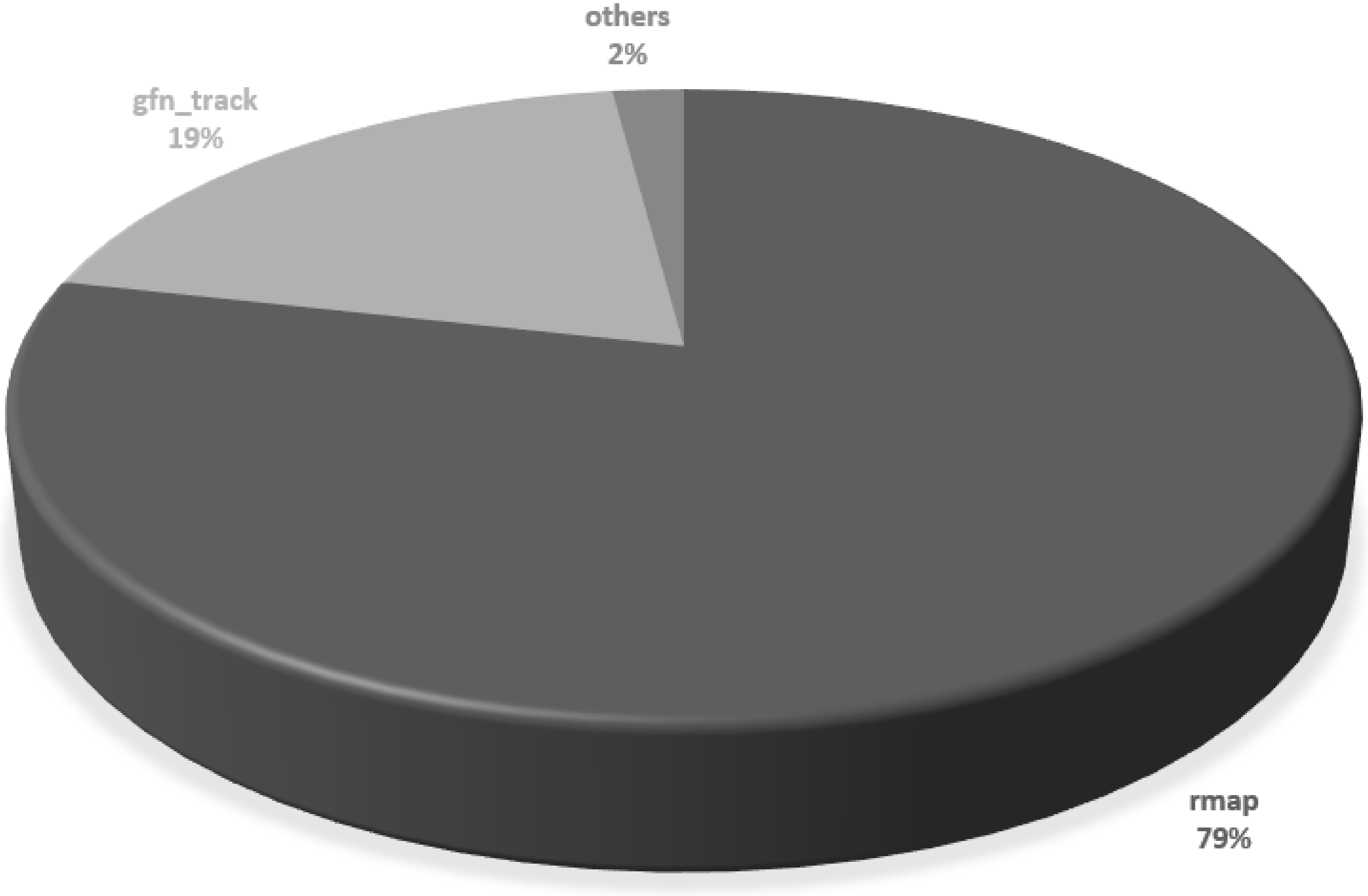
Agenda

- Background
- KVM memory cost proportion
- KVM guest page tracking optimization
- KVM rmap optimization
- Evaluation

Background

During the VM's life cycle, the KVM will use a lot of memory. Take Alibaba Cloud's large memory instance with 3Tib memory as an example. If 2M hugepage is enabled, the KVM uses nearly 8Gib of memory. For a large number of small-sized VMs, KVM will still consume a lot of memory, the used memory of KVM is basically linear with the memory sold. The issue must to be addressed.

KVM memory cost proportion



FROM THE ALIBABA CLOUD TEST ENVIRONMENT

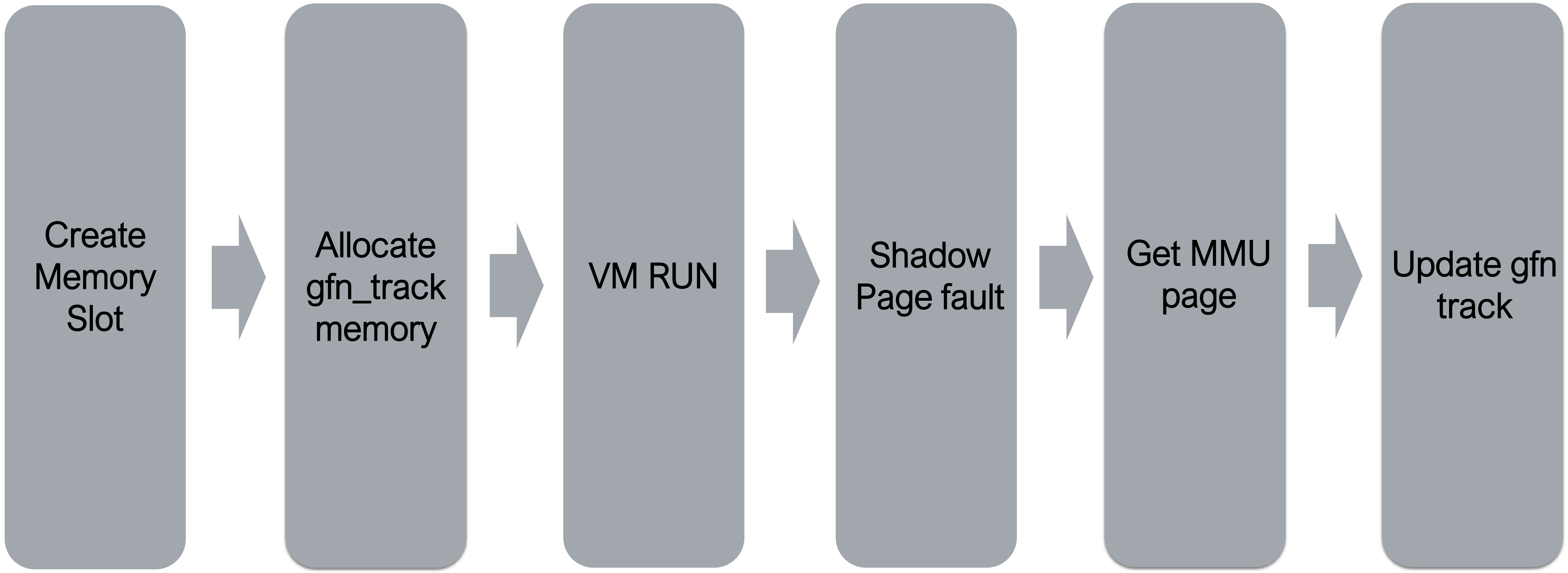
KVM guest page tracking optimization

What is gfn_track used for?

gfn_track is used to track page access in guest with shadow page table, only write access is tracked currently. The number of the accessed page will be recorded to page track bitmap.

But page track feature is not necessary for EPT VM now, so memory can be allocated until it is actually used.

KVM guest page tracking workflow



Create Memory Slot

Allocate gfn_track memory

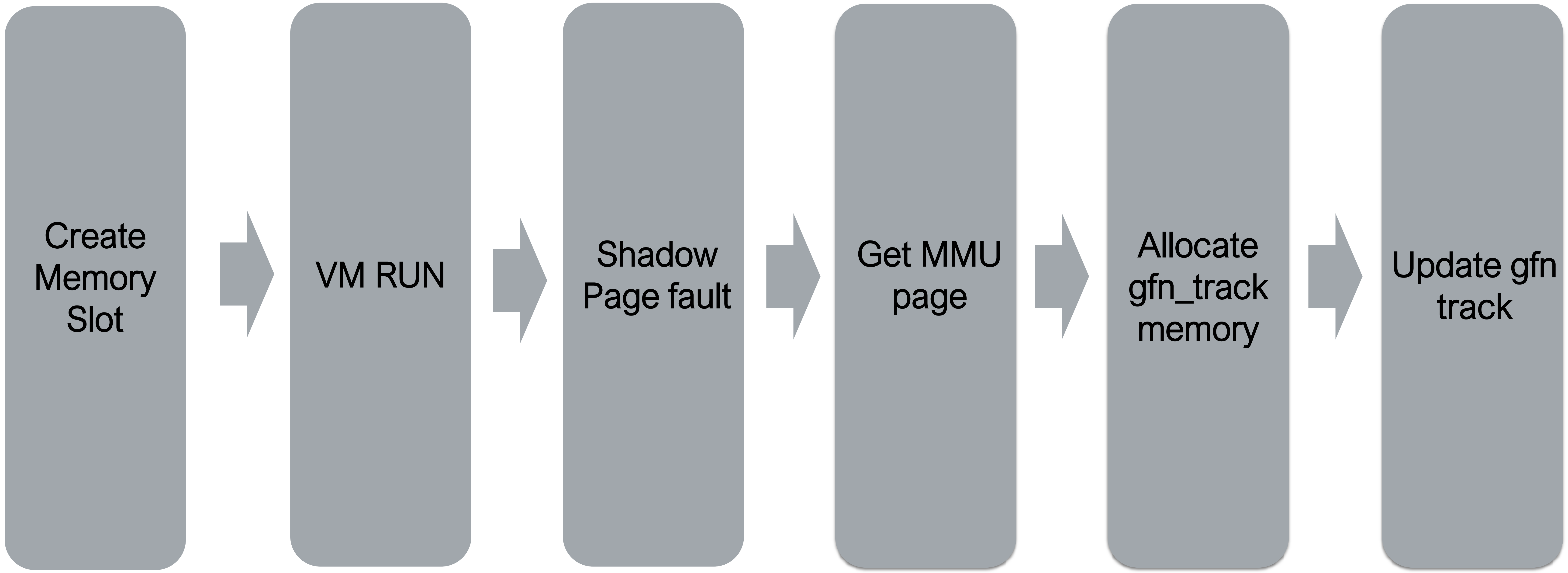
VM RUN

Shadow Page fault

Get MMU page

Update gfn track

KVM guest page tracking workflow optimization



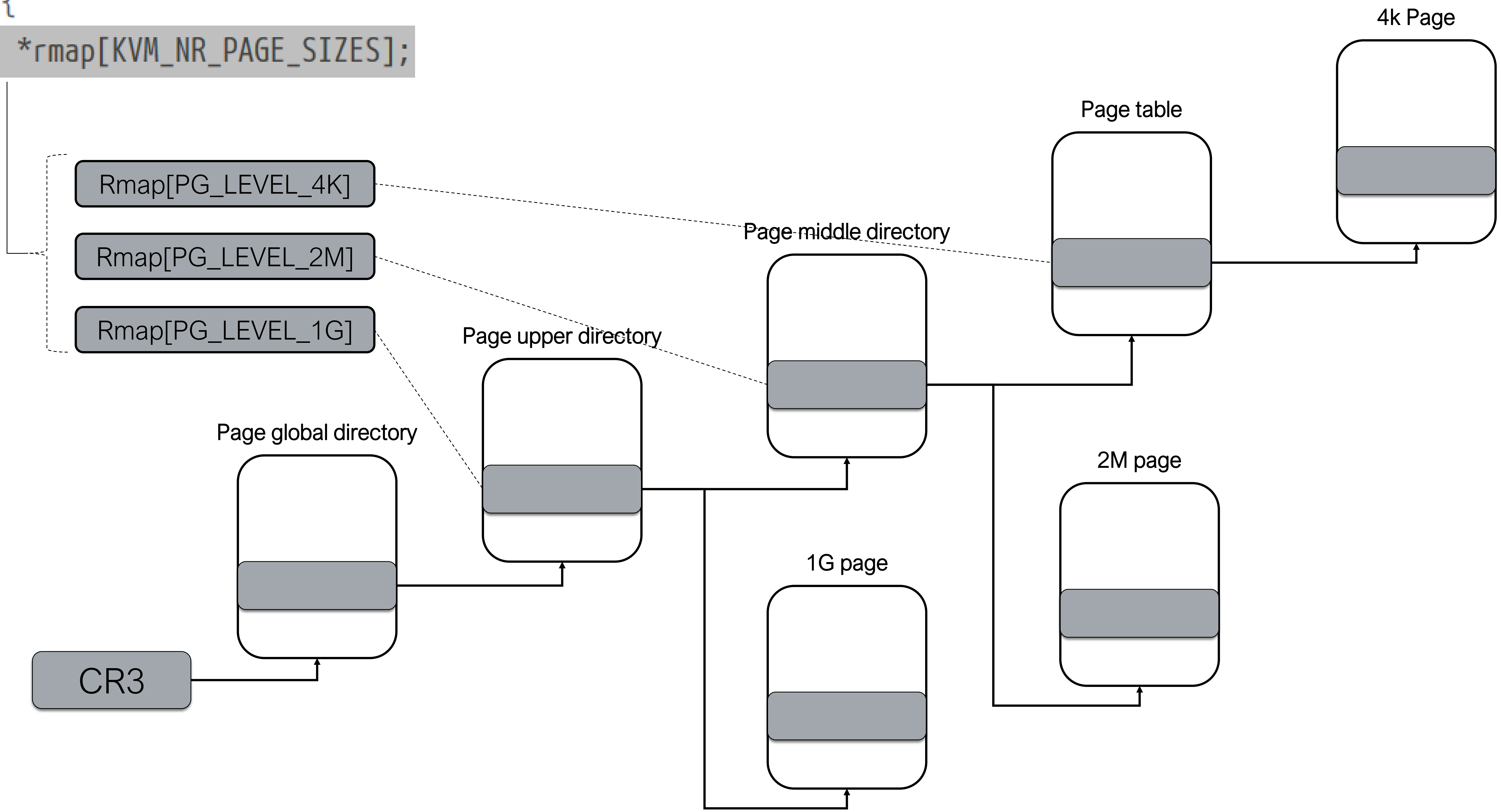
Create Memory Slot → VM RUN → Shadow Page fault → Get MMU page → Allocate gfn_track memory → Update gfn track

KVM rmap optimization

kvm rmap is widely used in KVM TDP MMU to accelerate the access of spte.

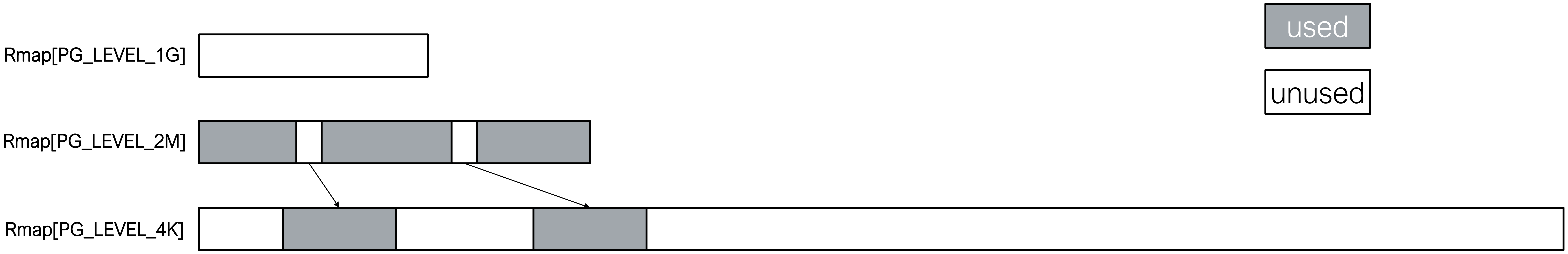
KVM rmap structure

```
struct kvm_arch_memory_slot {
        struct kvm_rmap_head *rmap[KVM_NR_PAGE_SIZES];
```



KVM rmap structure

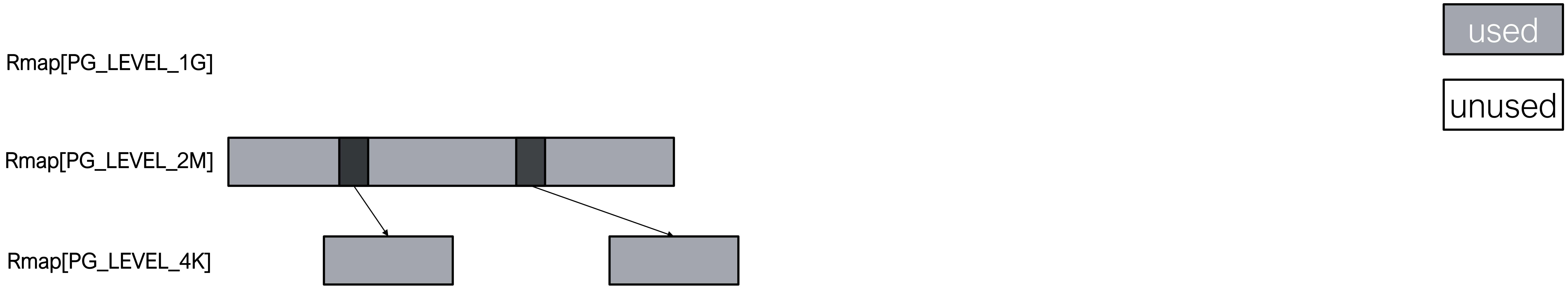
VM with 2M hugepage:



The reason why pg_level_4k is used is because some MMIOs are not aligned to 2M and drop to 4k pages

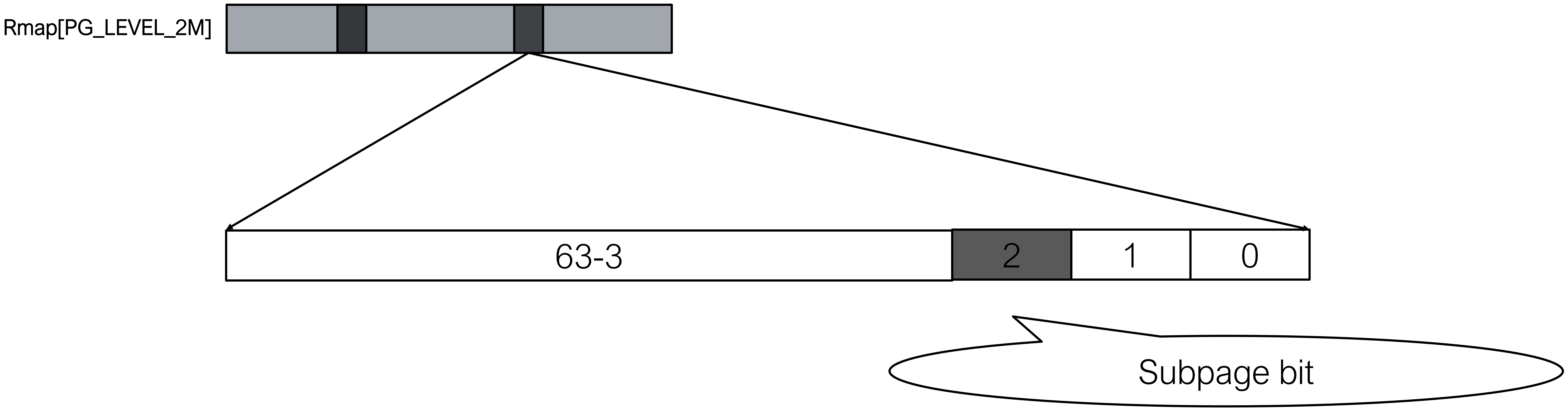
KVM rmap structure

VM with 2M hugepage:



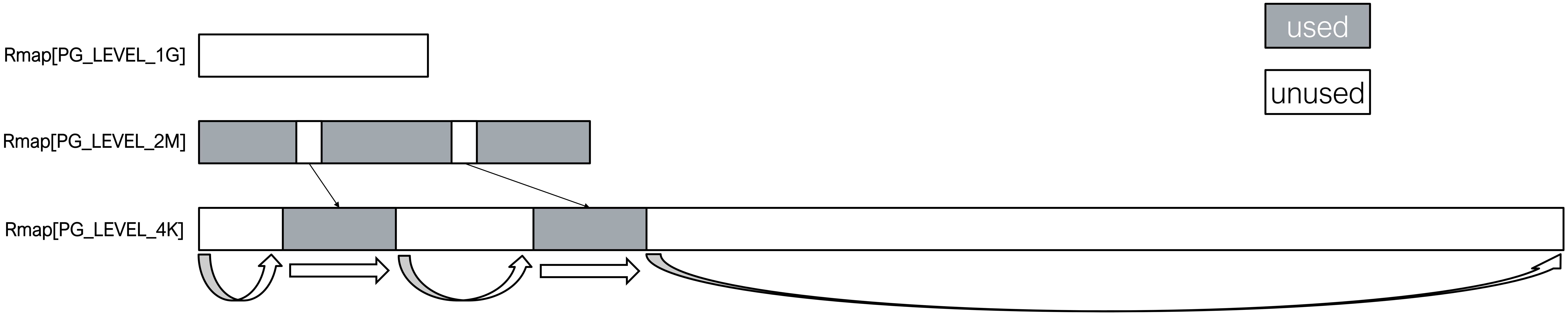
KVM rmap structure

How do I distinguish a subpage element from a SPTE element



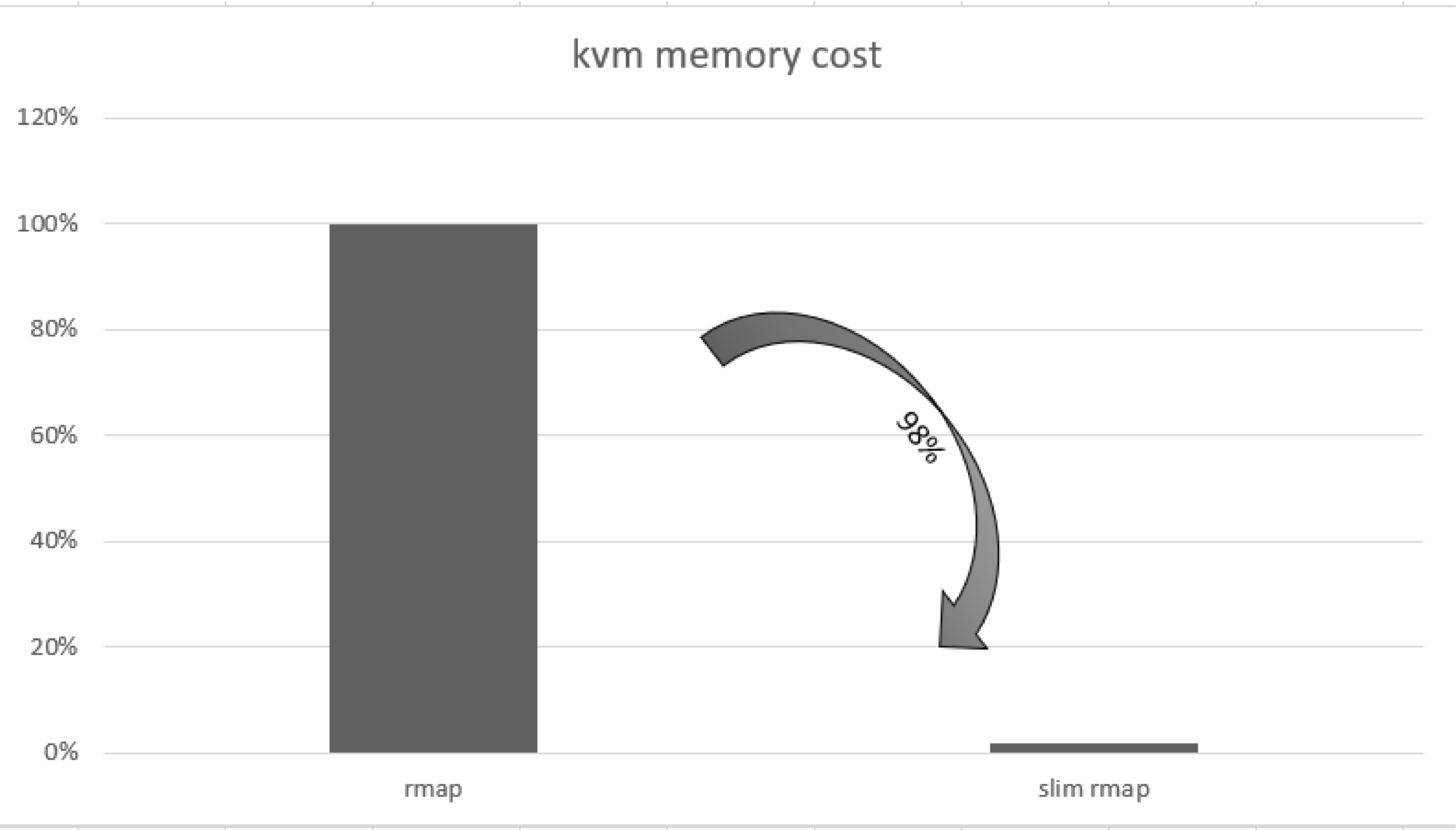
KVM rmap structure

How to traverse sptes:



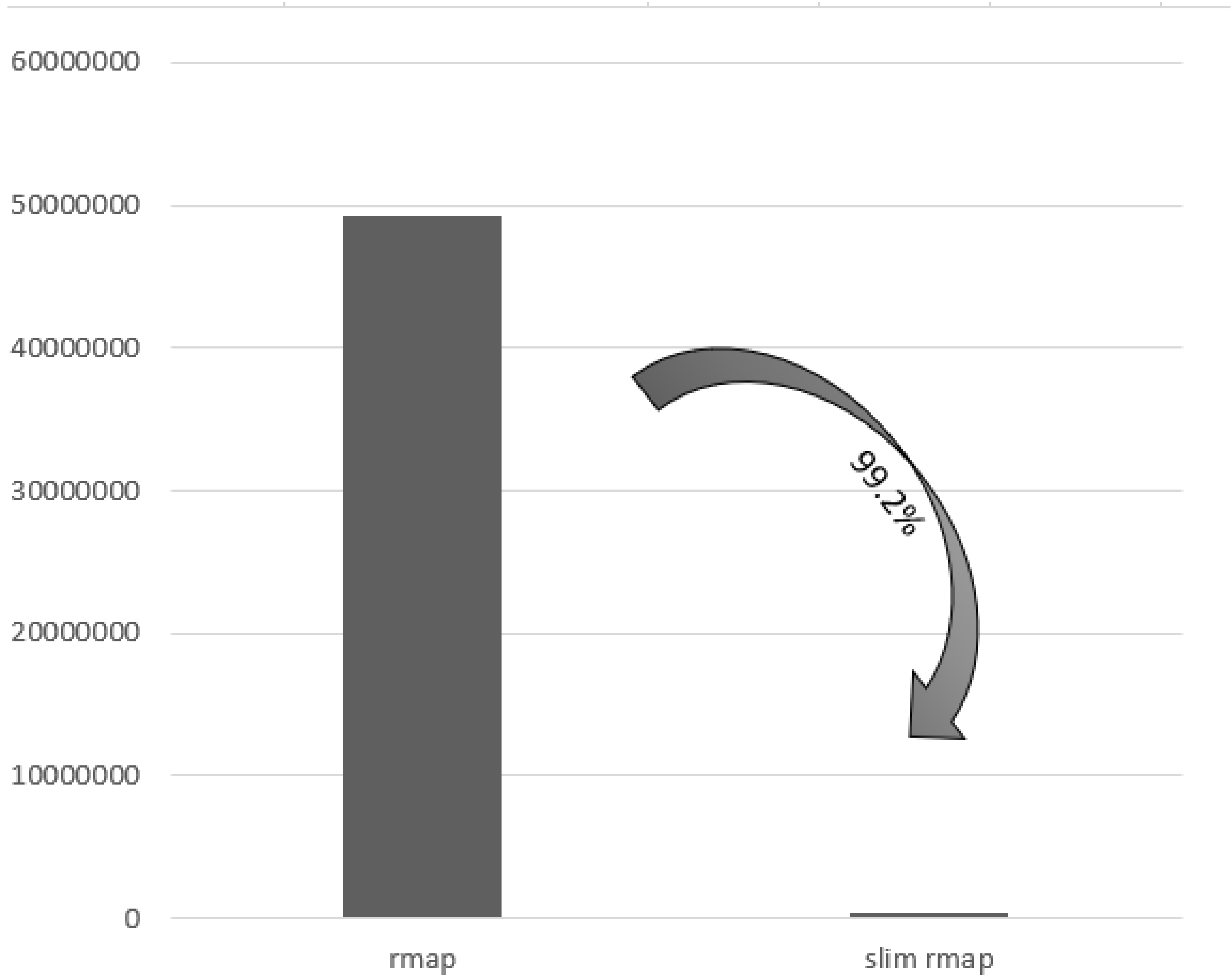
Evaluation

The KVM memory cost of VM with 2M hugepage



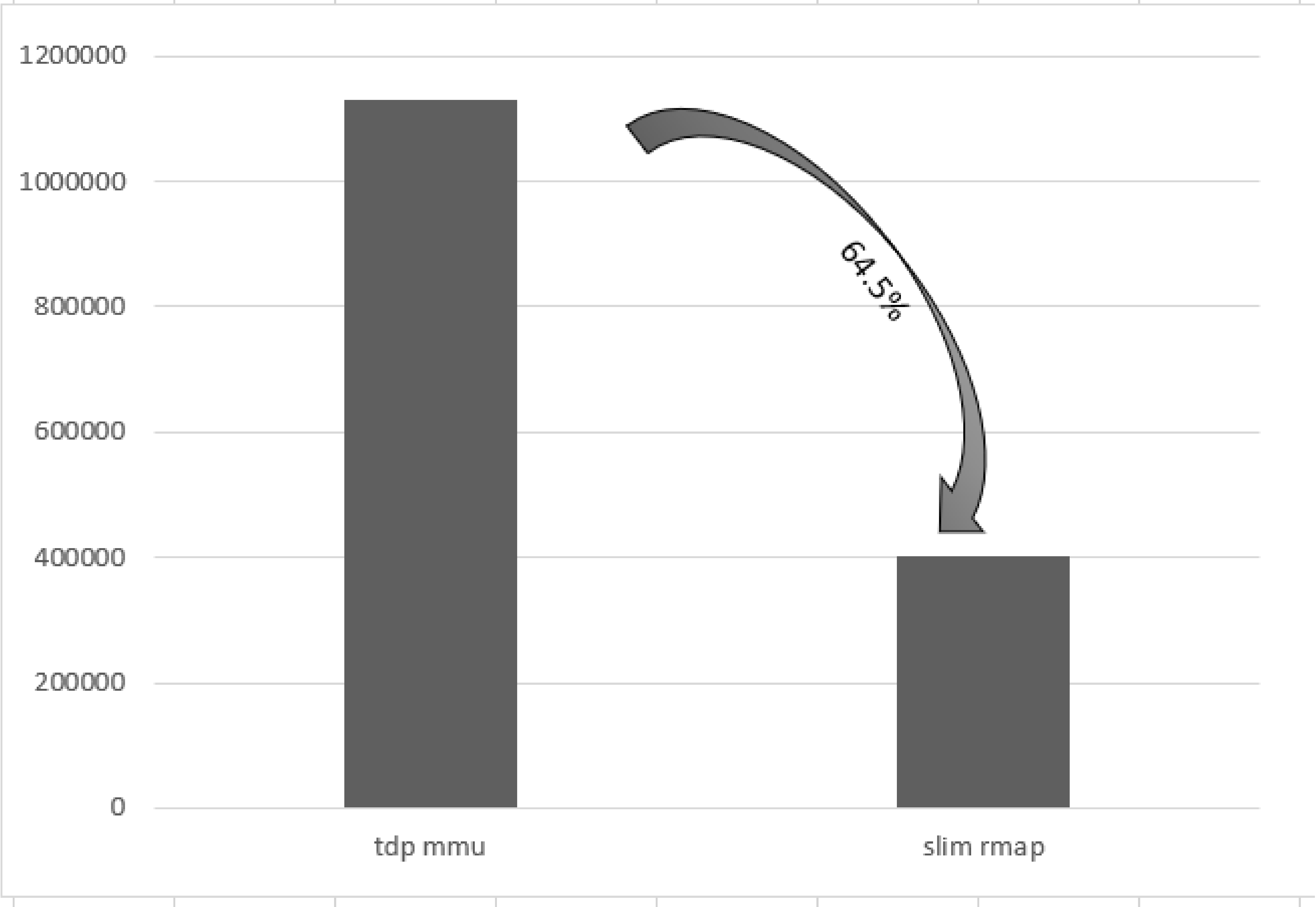
Evaluation

The time to traverse all sptes from a 32G VM



Evaluation

The TDP MMU feature has removed rmap in upstream, here is the time comparison of the two features traverse all sptes from a 32G VM.



Evaluation

The current problem is that when the VM is migrated using PML, all 2M hugepages will be dropped into 4K pages, which will increase the kvm memory cost, but when the migration is completed or fails, this part of the memory will be freed. and for the one by one migration in host, the issue looks acceptable.

Q&A

Alibaba Cloud | MORE THAN JUST CLOUD