Alibaba Cloud

Worldwide Cloud Services Partner

Speed Up Boot-up Time for Guest in Alibaba Cloud

Li Weinan william.lwn@alibaba-inc.com
Guo Cheng hanyu.gc@alibaba-inc.com

Agenda

- Background
- Async dma map
- Guest boot process with async dma map
- Optimization design
- Achievements

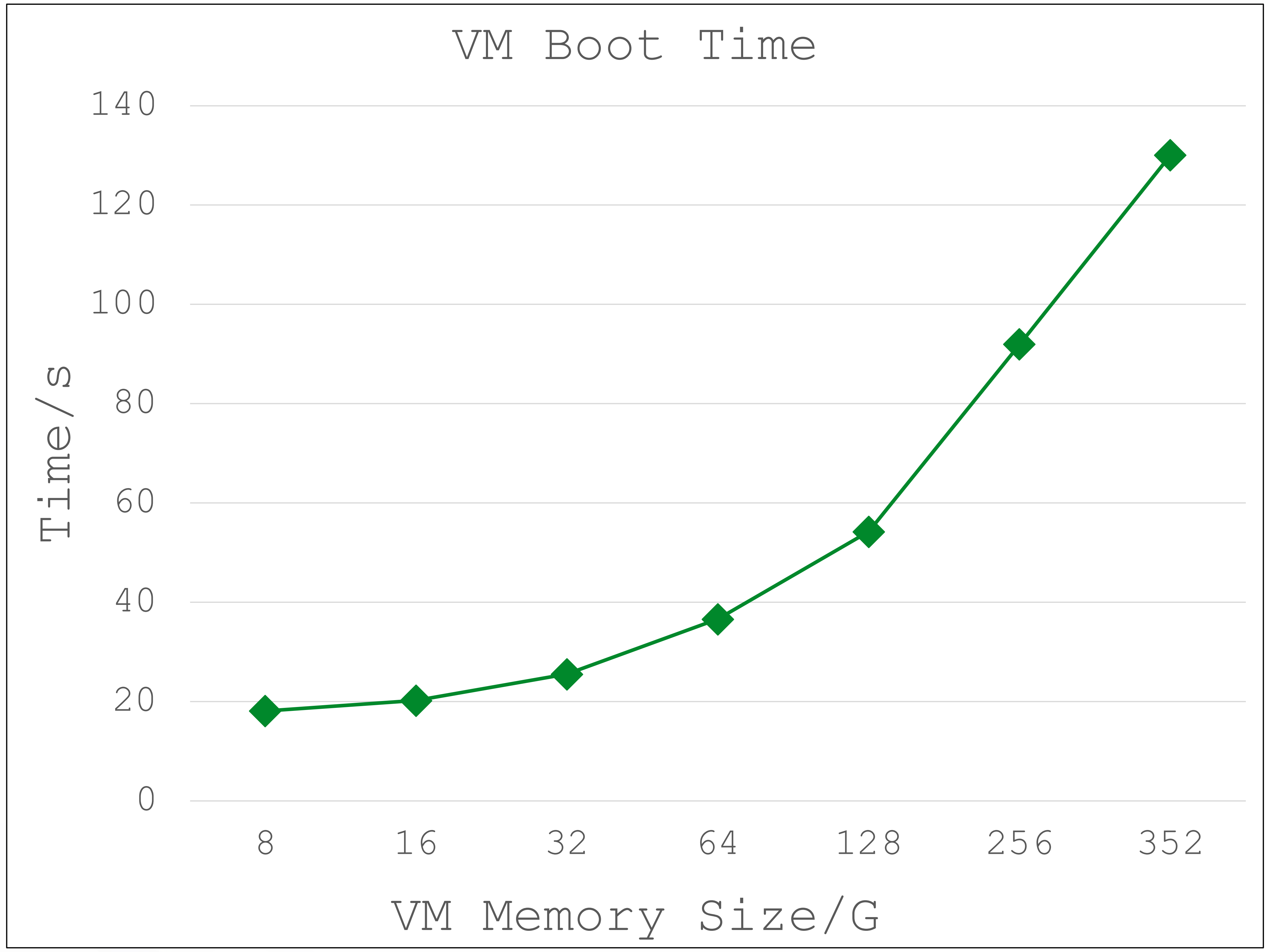
What is the problem?

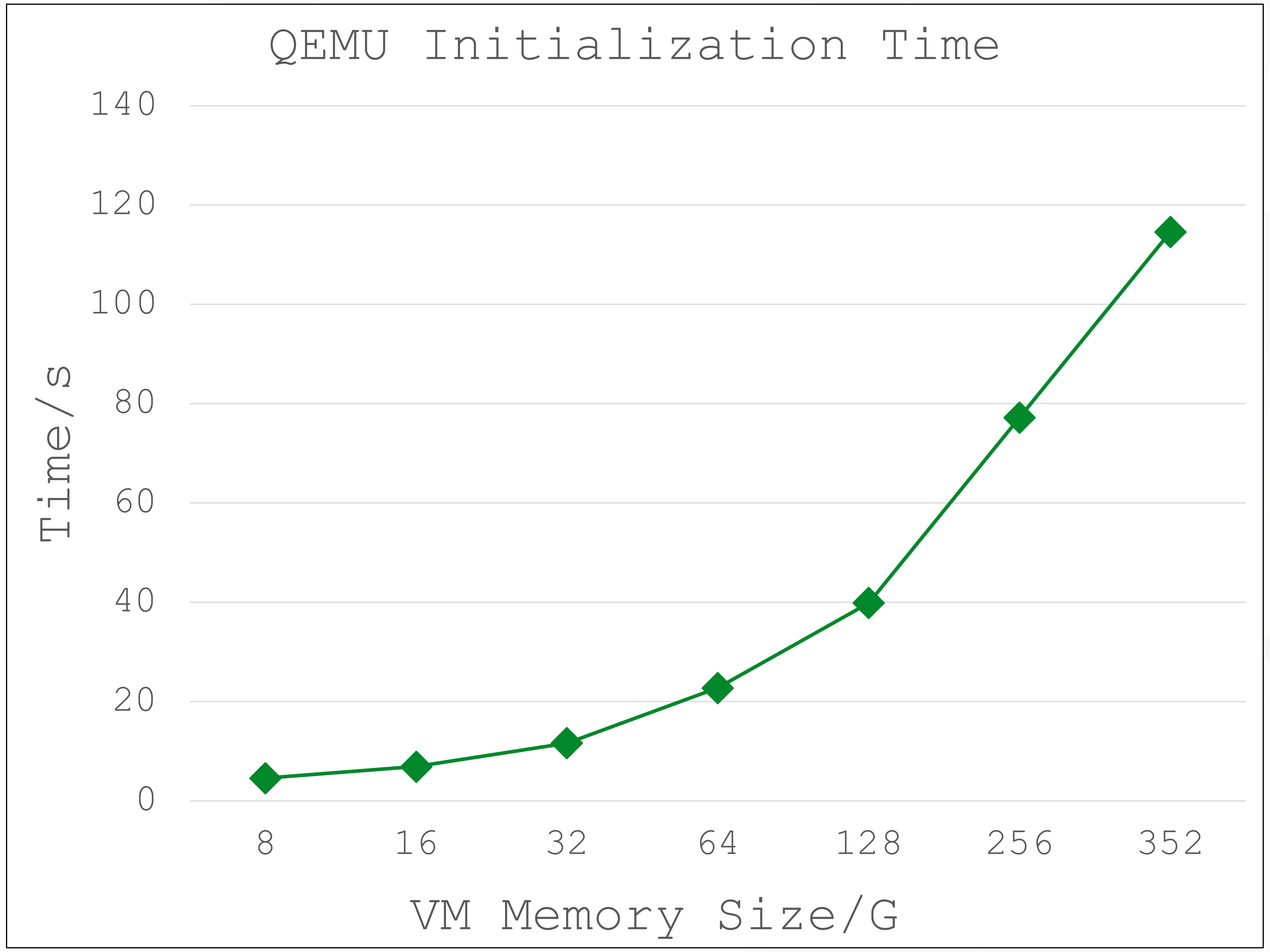
- Dma_map all the guest memory when there is passthrough device

- 8G->384G

  Dma_map time is one big problem!

Guest boot & QEMU initialization time vs memory size





Conditions

- More time costs along with more memory
- No DMA no dma_map
- DMA specific range memory

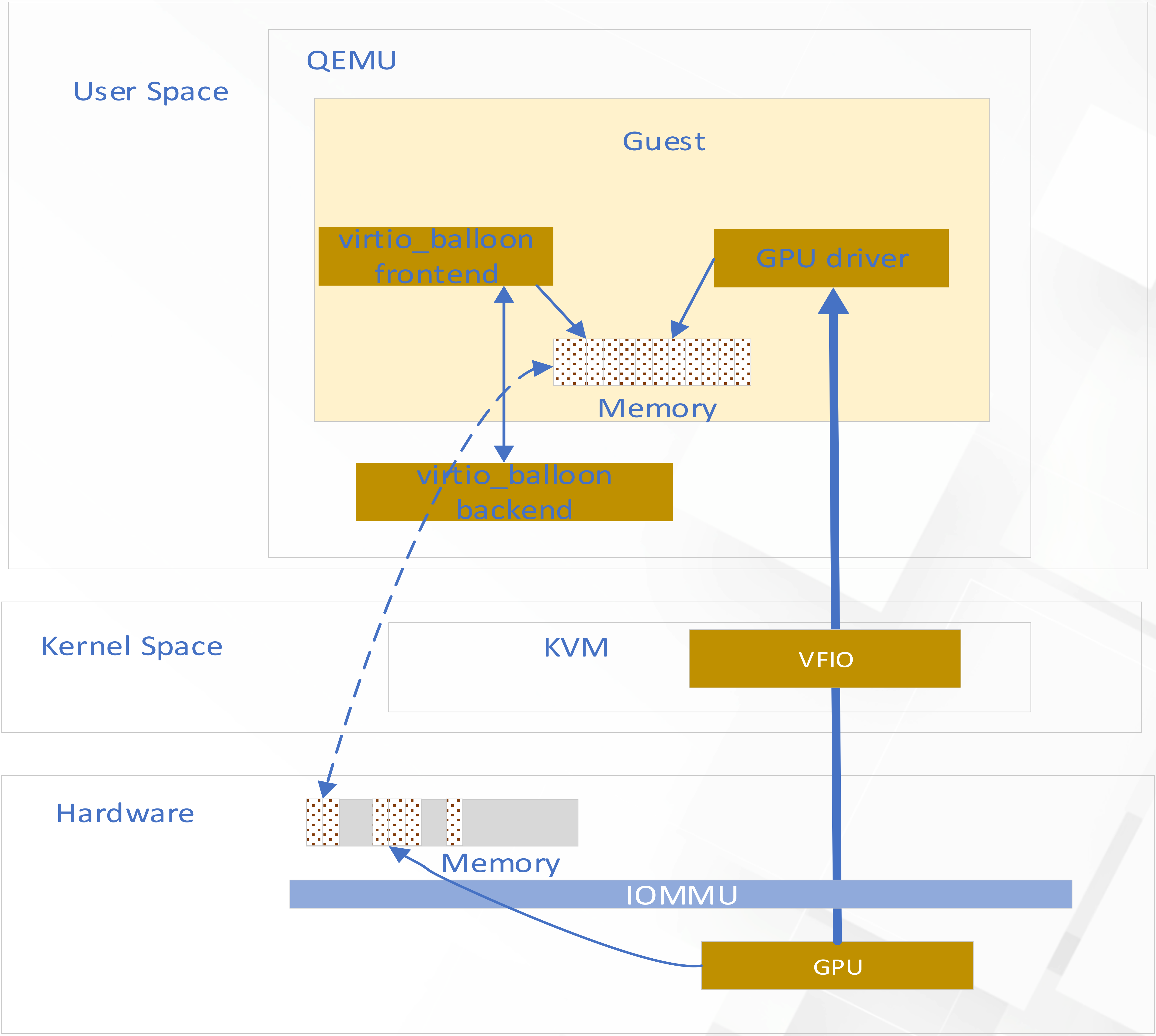
Options

- virtual IOMMU

- Async dma map
  - Only map necessary memory first
  - Map asynchronously in the background

Async dma_map

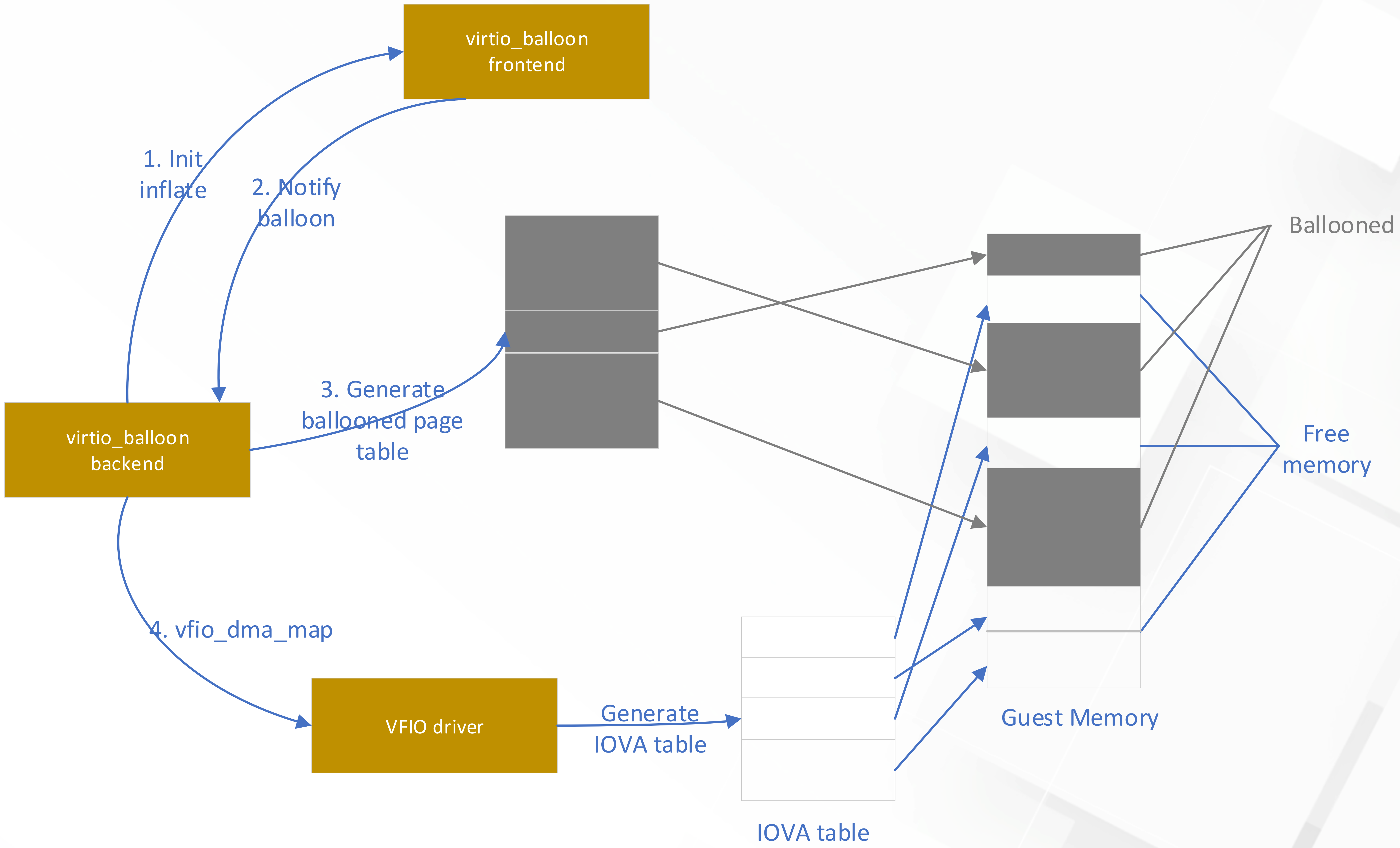
✭Balloon memory before allocated for DMA



Overview of memory access with a passthrough device

Architecture Overview

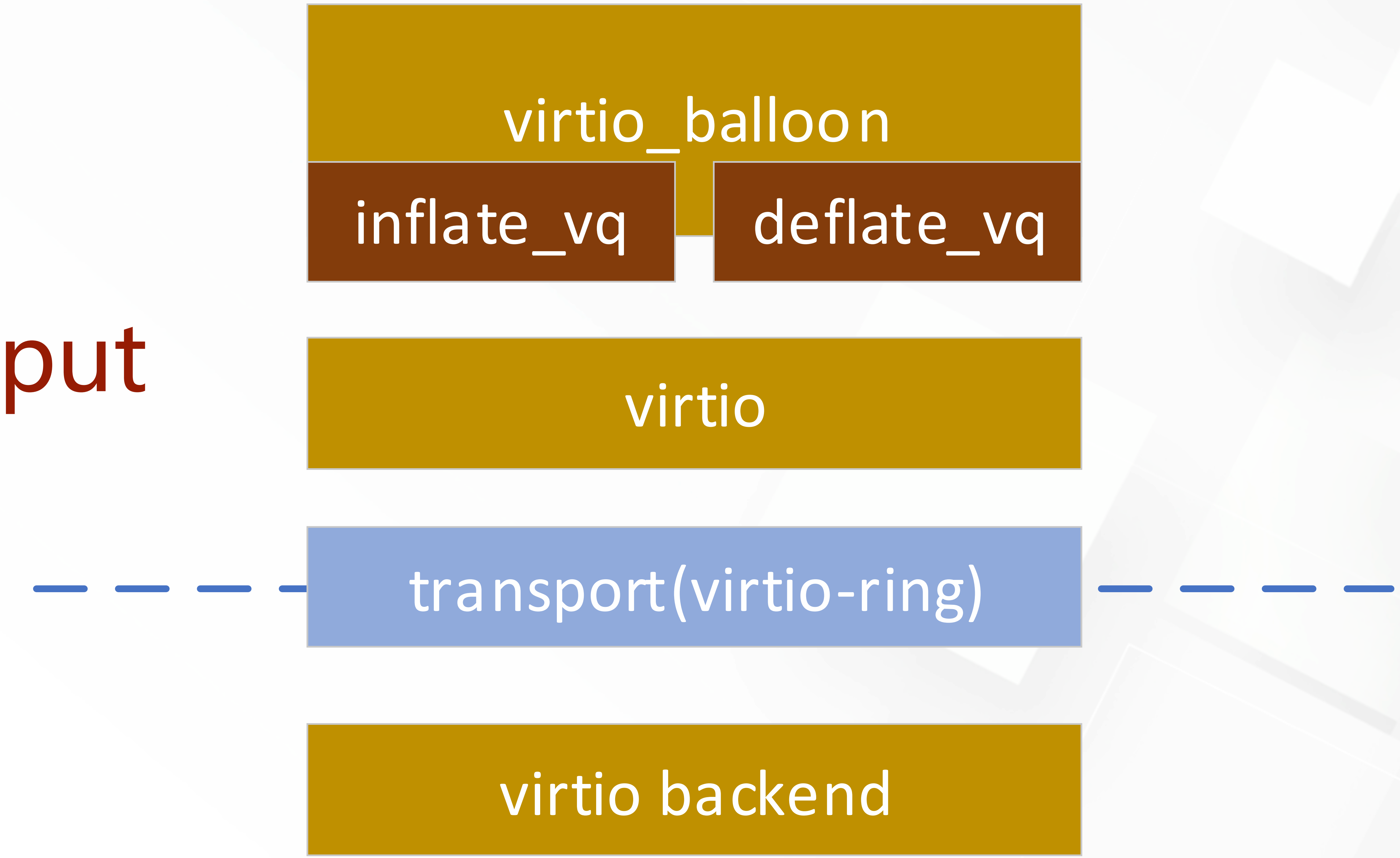
- ➢ QEMU
  - Trigger vfio_dma_map
  - Trigger balloon change
  - Track ballooned pages
- ➢ virtio_balloon driver
  - Balloon pages
  - Tell to host (QEMU)
- ➢ VFIO driver
  - do vfio_pin_map_dma



virtio_balloon communication

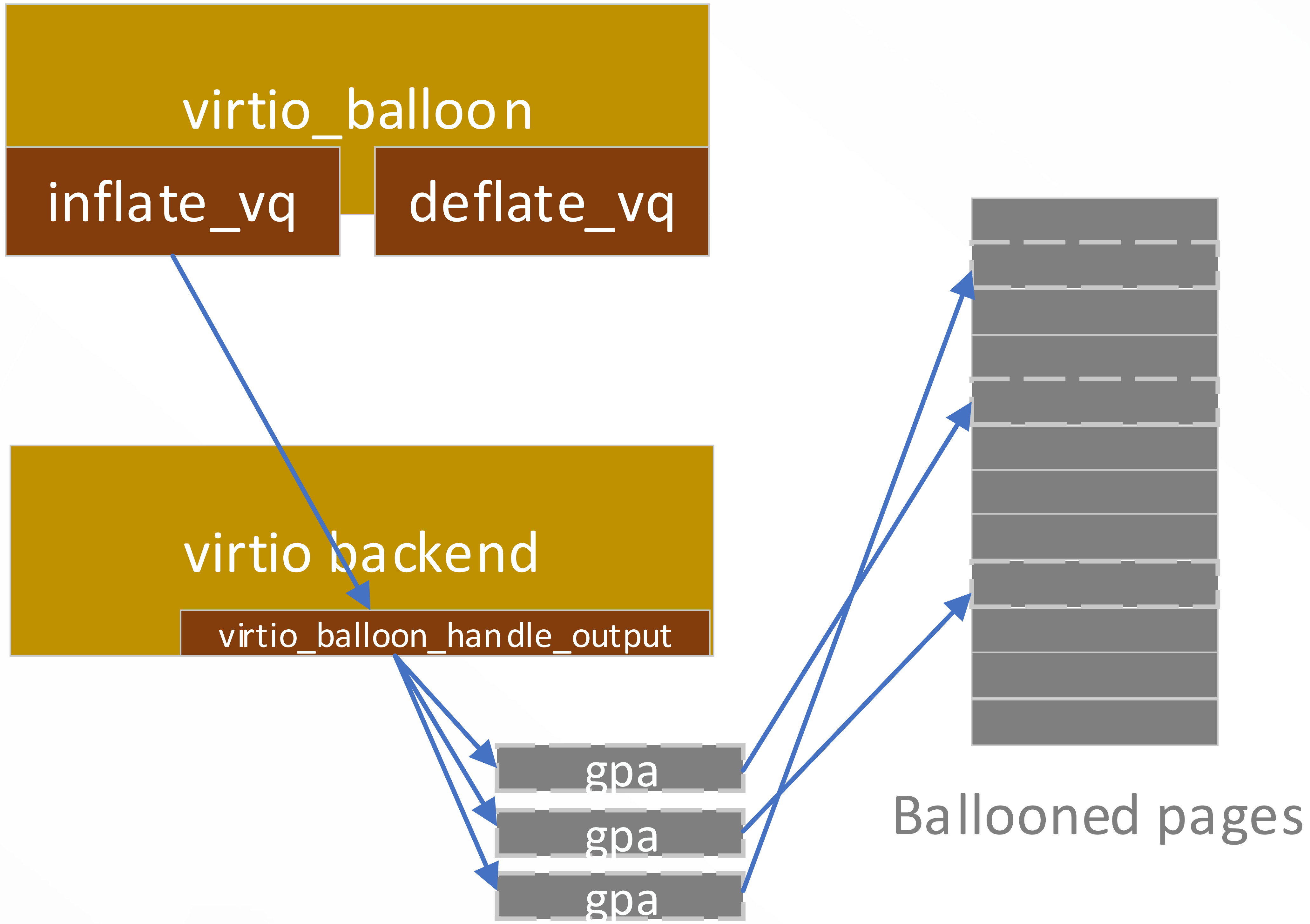
Related functions and struct

- inflate_vq
- deflate_vq
- virtio_balloon_handle_output
- VirtQueueElement:
  guest PFN
  page_num

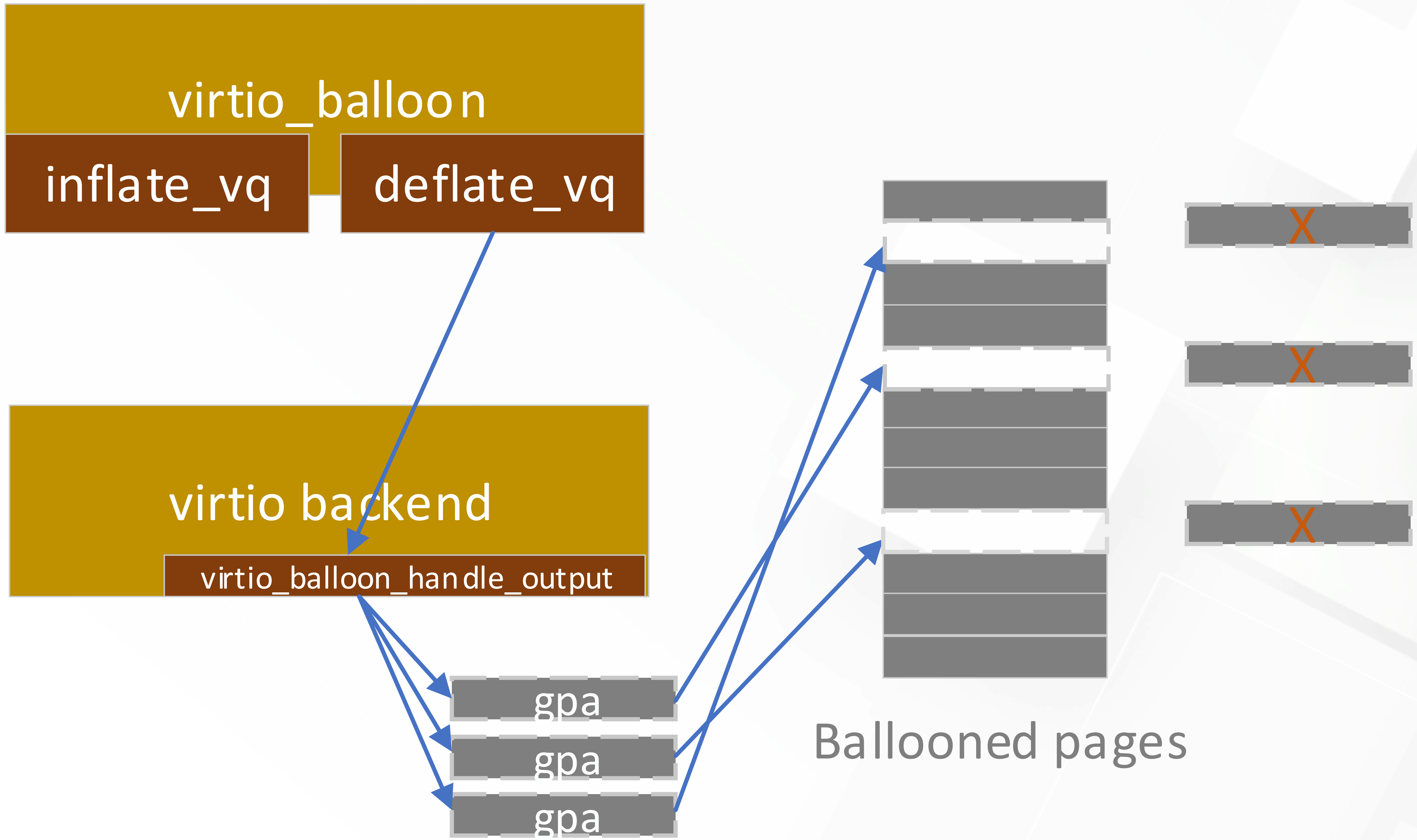


Balloon range tracking workflow

Alibaba Cloud

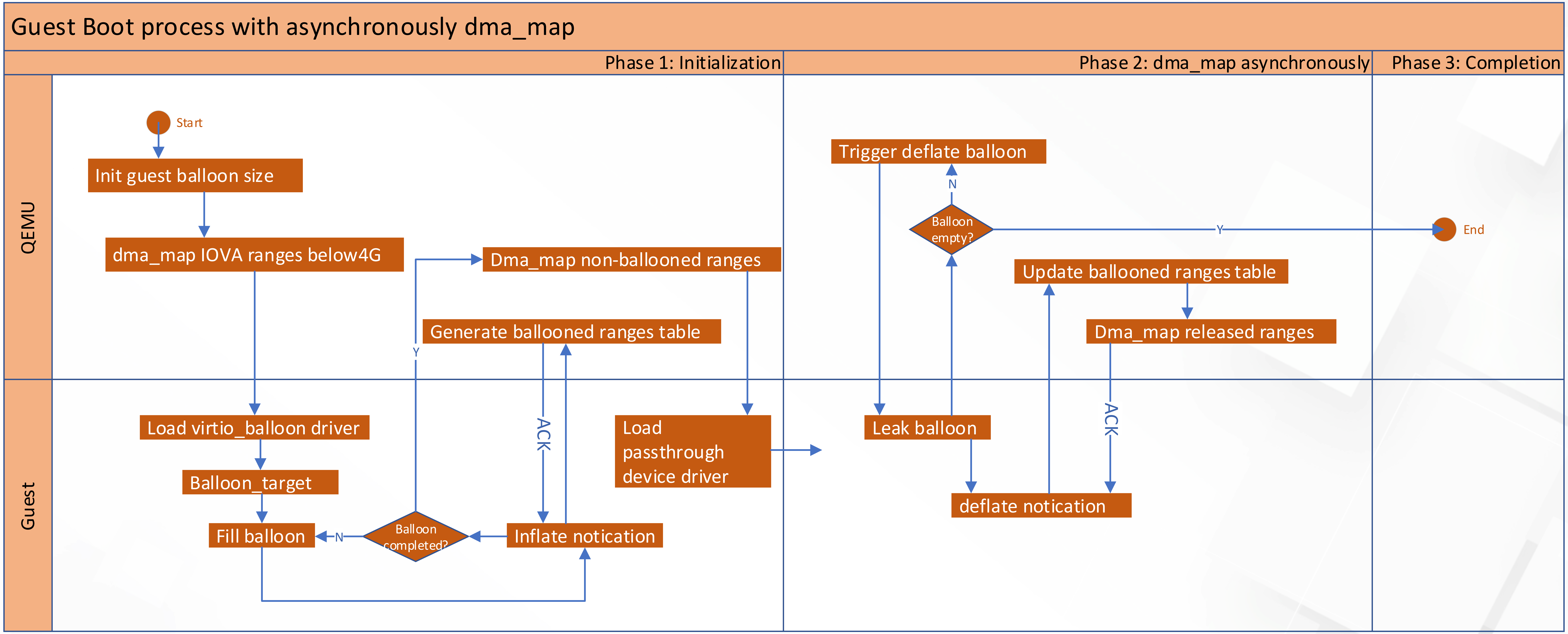


Inflate process



Deflate process

Guest boot process with async dma map


Guest Boot process with asynchronously dma_map
Phase 1: Initialization
Phase 2: dma_map asynchronously
Phase 3: Completion
QEMU
Guest
Start
Init guest balloon size
dma_map IOVA ranges below4G
Dma_map non-ballooned ranges
Generate ballooned ranges table
Y
Load virtio_balloon driver
Balloon_target
Fill balloon
N
Balloon completed?
Inflate notication
ACK
Load passthrough device driver
Trigger deflate balloon
N
Balloon empty?
Y
End
Update ballooned ranges table
Dma_map released ranges
Leak balloon
deflate notication
ACK

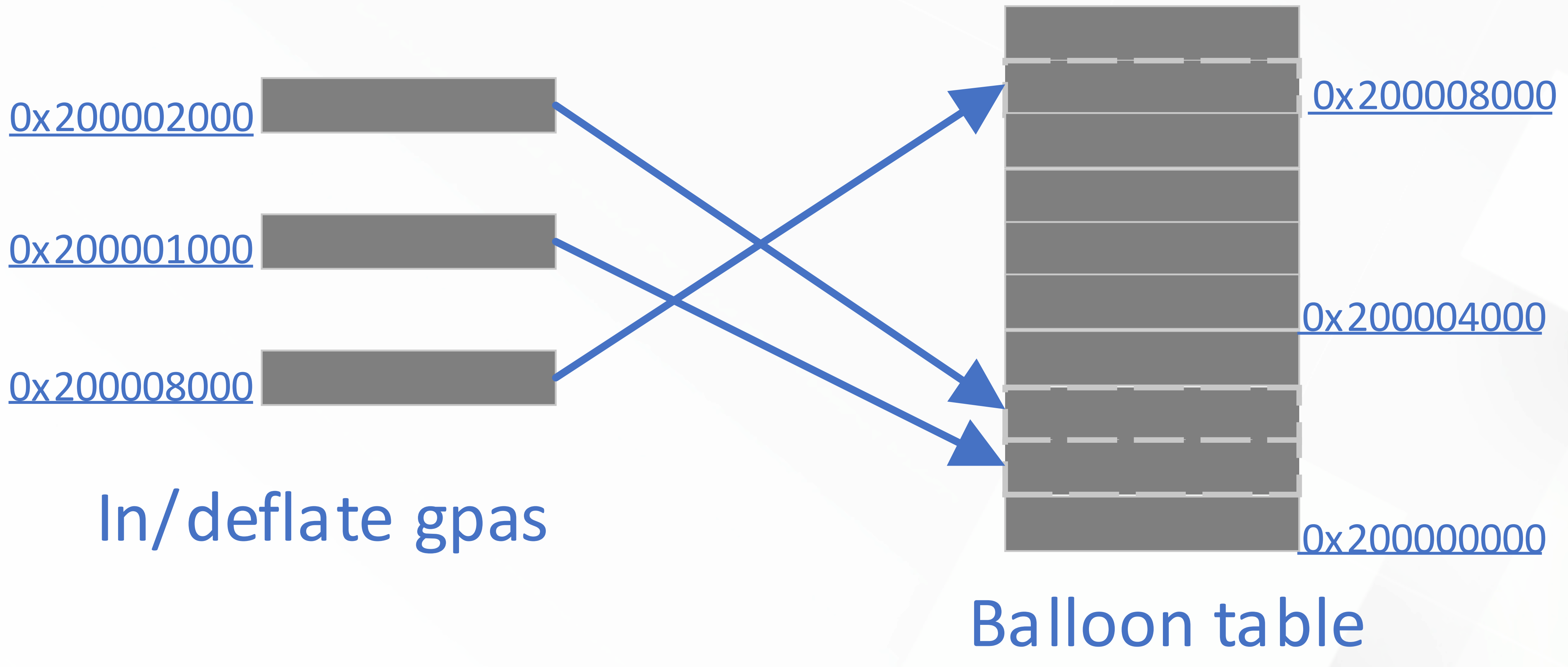
Optimization design

Auto-combination

Problem:

□ one page at a time

□ 256 pages per cycle



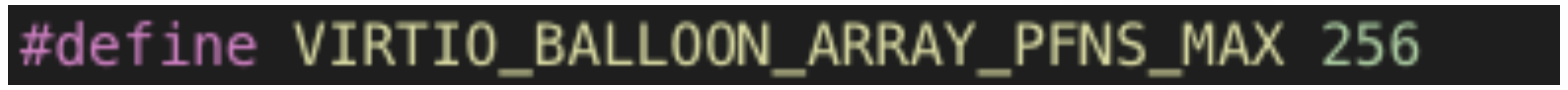
Practice:

✓ **Combine adjacent pages**

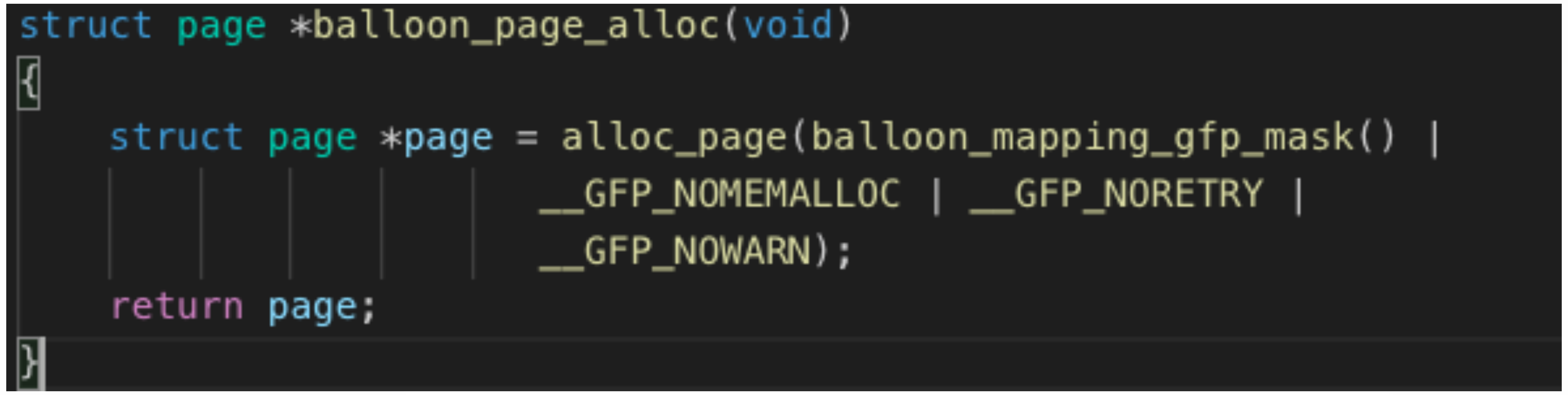
- Most of the memory ranges are adjacent
- dma_map after inflate balloon process finished

Optimization design

Increase balloon page size



```
#define VIRTIO_BALLOON_ARRAY_PFNS_MAX 256
```



```
struct page *balloon_page_alloc(void)
{
    struct page *page = alloc_page(balloon_mapping_gfp_mask() |
                                   __GFP_NOMEMALLOC | __GFP_NORETRY |
                                   __GFP_NOWARN);
    return page;
}
```

□ 4K page is too small which will import heavy but unnecessary communication between guest and host

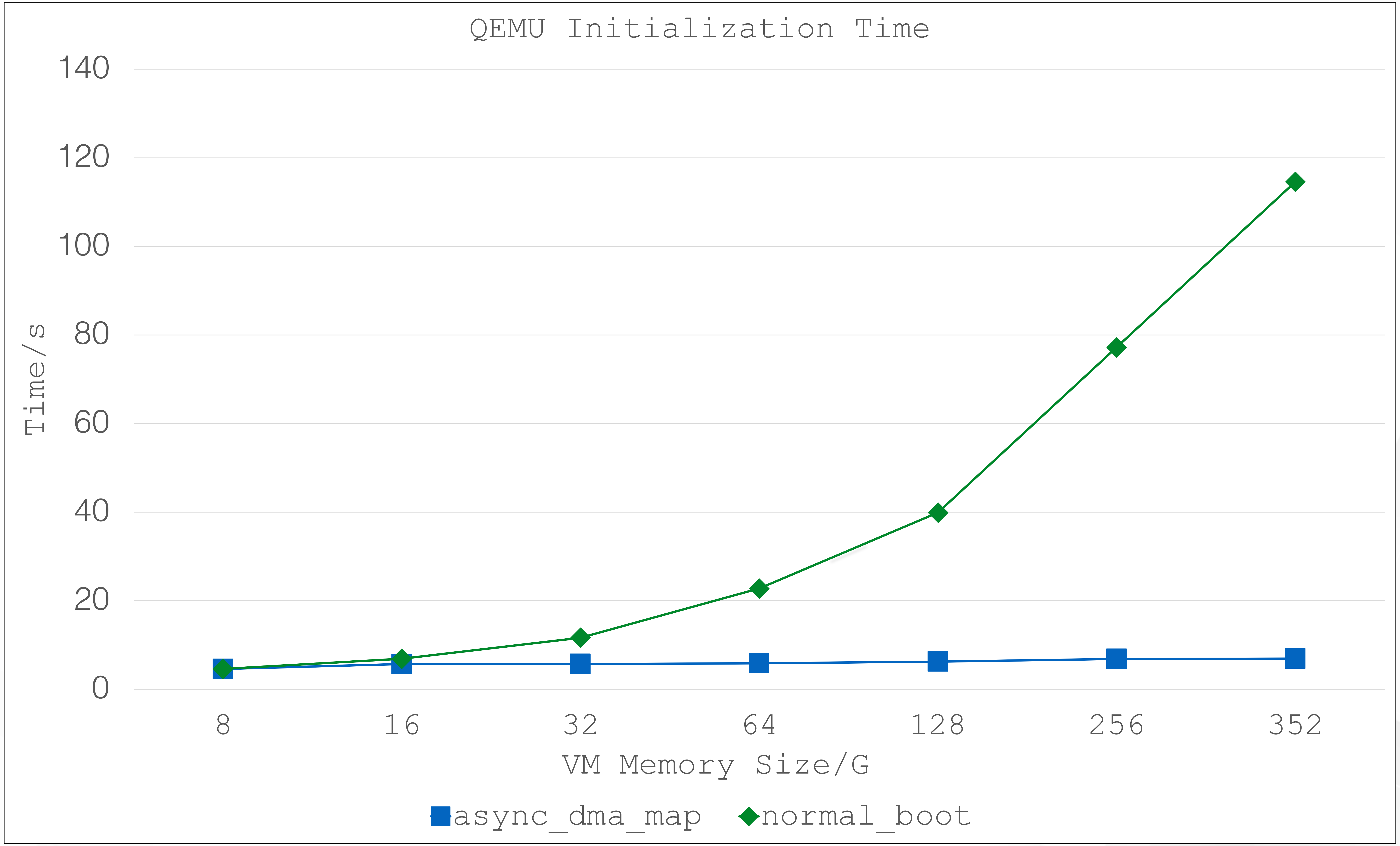
✓ 4KB -> 2MB, one virtio talk can in/deflate 2 x 256 = 512MB memory

Optimization design

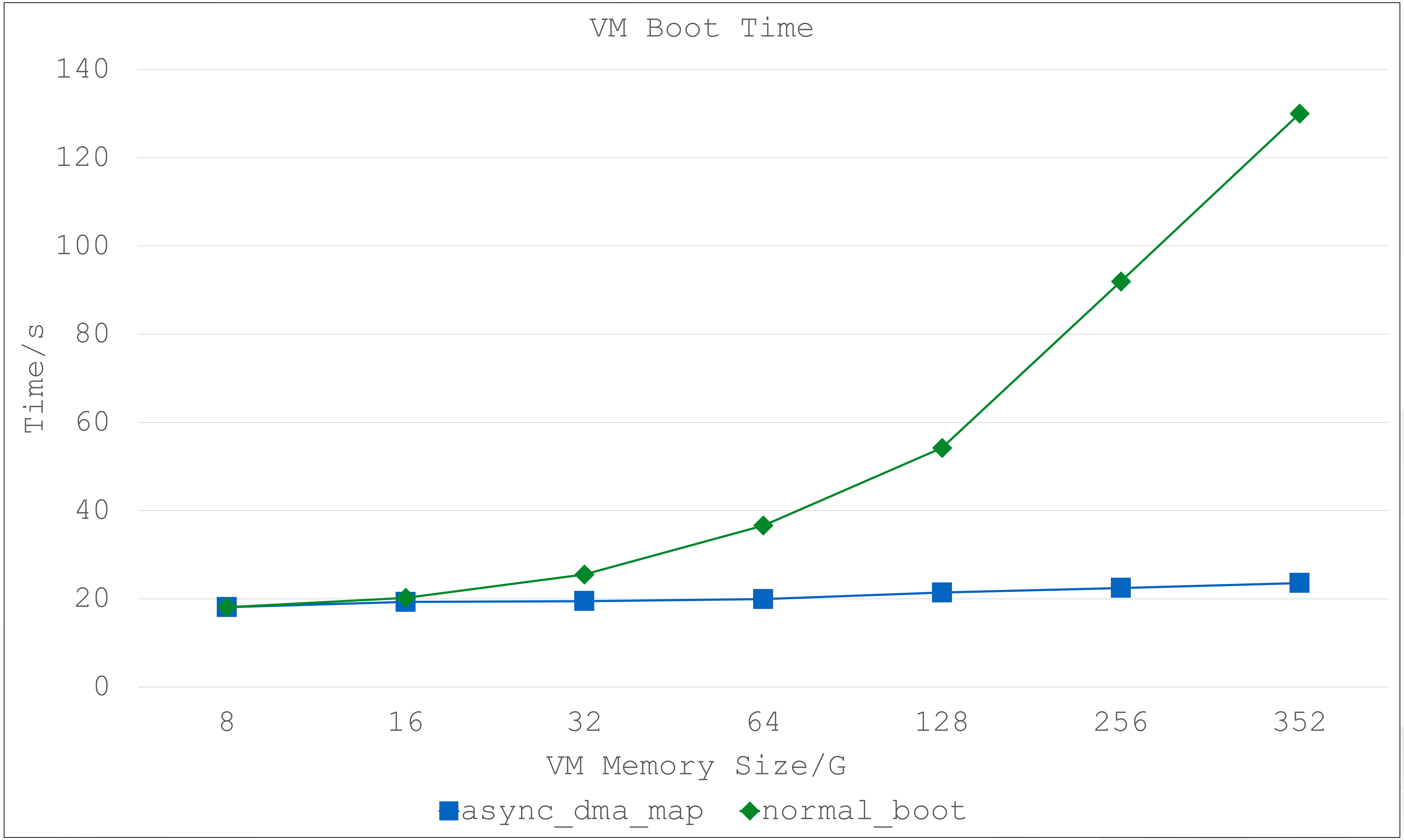
✓ Pre-map to perform dma_map

- Asynchronously dma_map can start early independent of deflating notification
- Insert new dma_map range if the released pages beyond mapped ranges

Achievements



Achievements



Q&A

Li Weinan
william.lwn@alibaba-inc.com

Guo Cheng
hanyu.gc@alibaba-inc.com