In kernel memory compression

Bob Liu
bob.liu@oracle.com

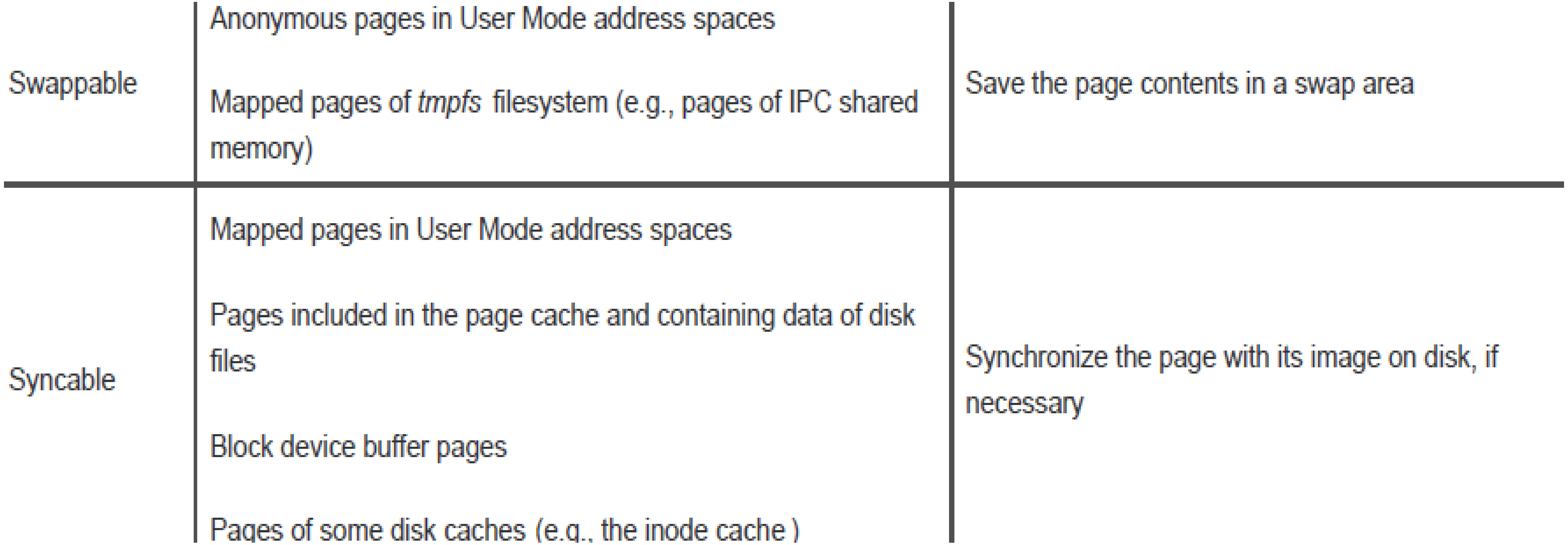
Oct 2013

Overview

- Why memory compression
- Zswap introduction
- Zswap usage and performance
- Zswap challenge
- Zram
- Zcache

What happens if system is under memory pressure

- **Linux kernel will do page reclaim**
- **Typical pages to be considered:**

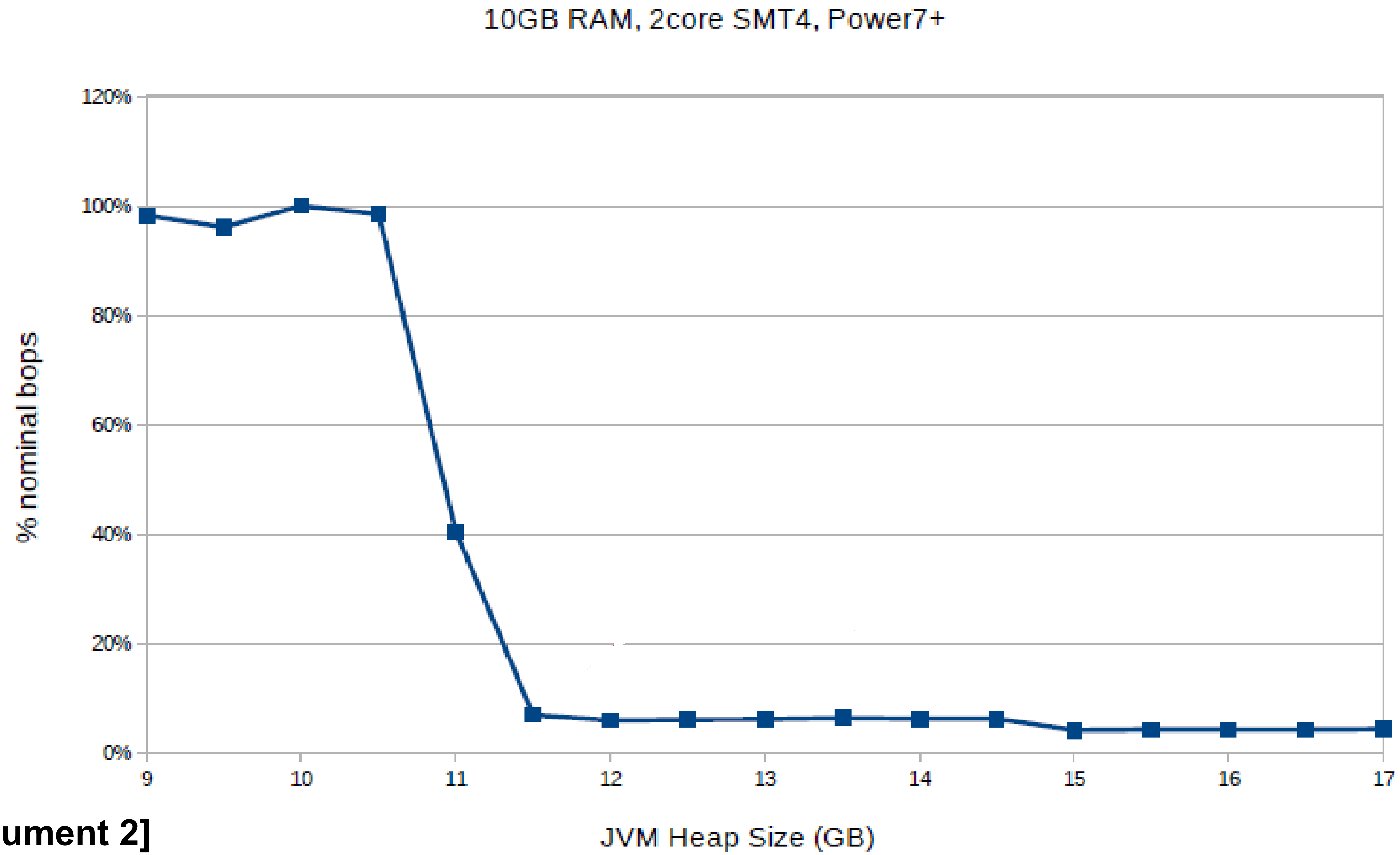


Swappable	Anonymous pages in User Mode address spaces Mapped pages of *tmpfs* filesystem (e.g., pages of IPC shared memory)	Save the page contents in a swap area
Syncable	Mapped pages in User Mode address spaces Pages included in the page cache and containing data of disk files Block device buffer pages Pages of some disk caches (e.g., the inode cache)	Synchronize the page with its image on disk, if necessary

Consider swappable pages only

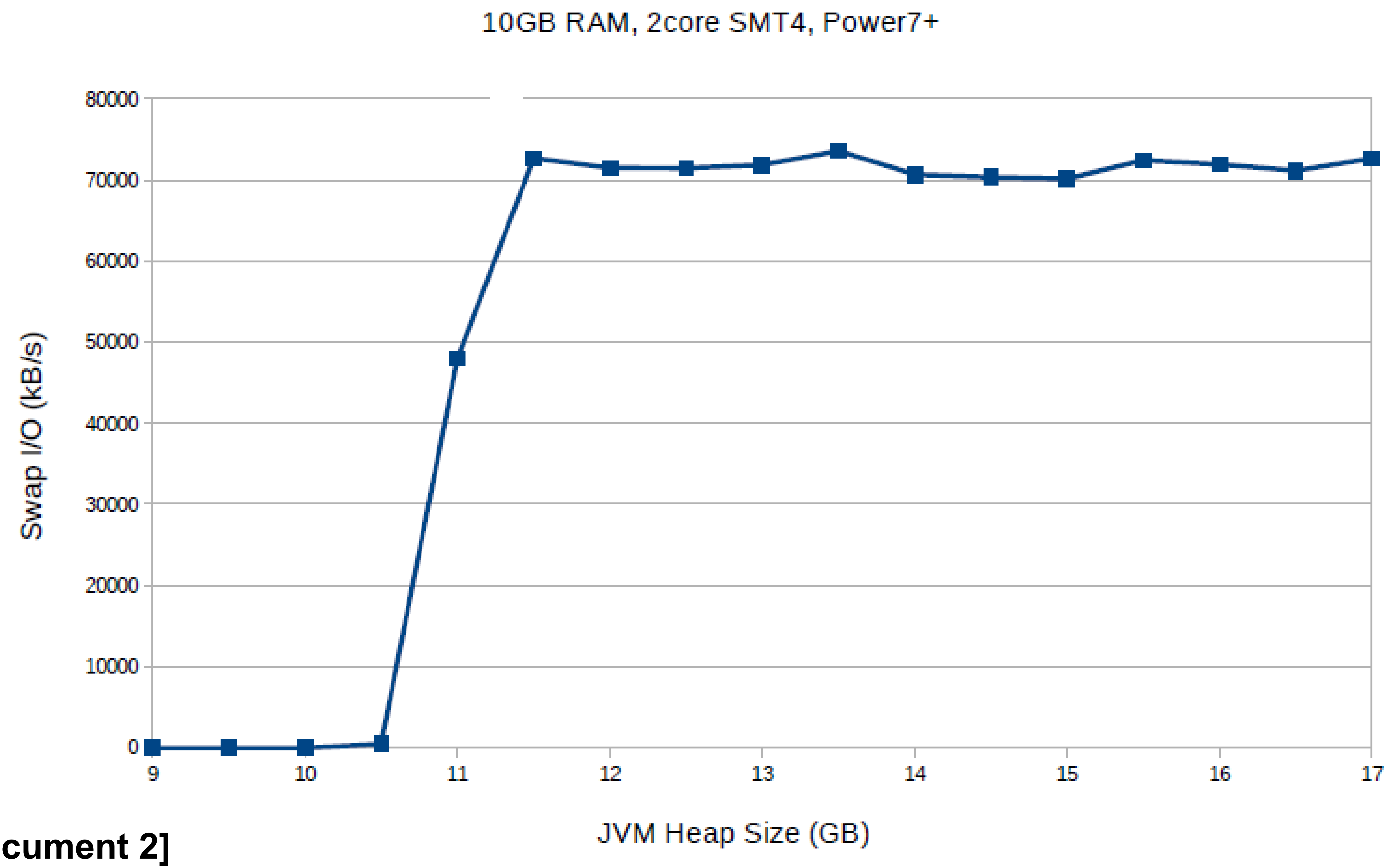
- Pages will be swapped out to disk
- Disk is much slower
- Swap will cause significant performance drop

SPECjbb Performance



* [Reference document 2]

Swap I/O



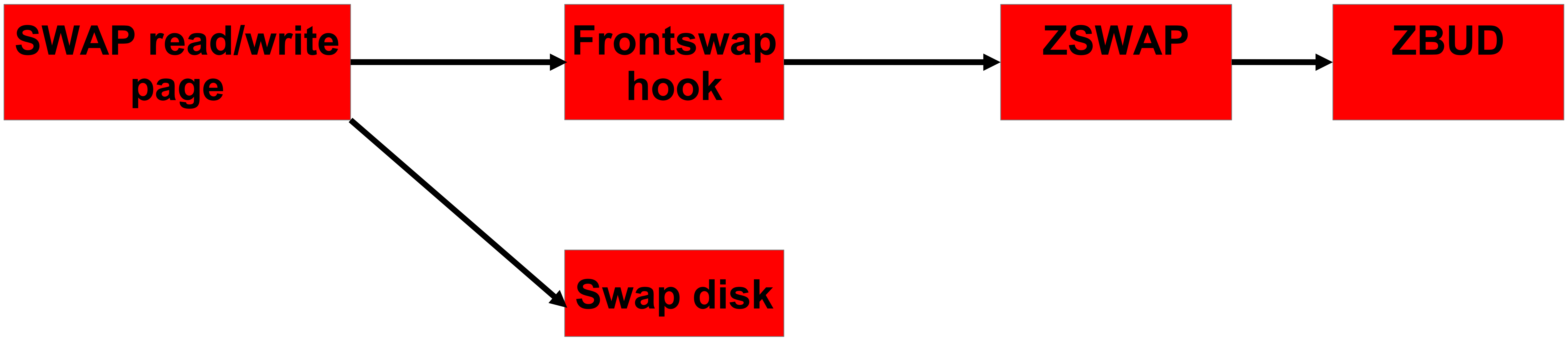
* [Reference document 2]

A way to smooth out this I/O storm and performance

ZSWAP

- Hook into swap_writepage() via the frontswap API
- Compress the swap page
- Store the compressed page in a dynamically allocated memory pool(ZBUD)
- Hook into swap_readpage()
- Decompress from memory pool
- Avoid touching the slow swap device

ZSWAP



SWAP read/write page

Frontswap hook

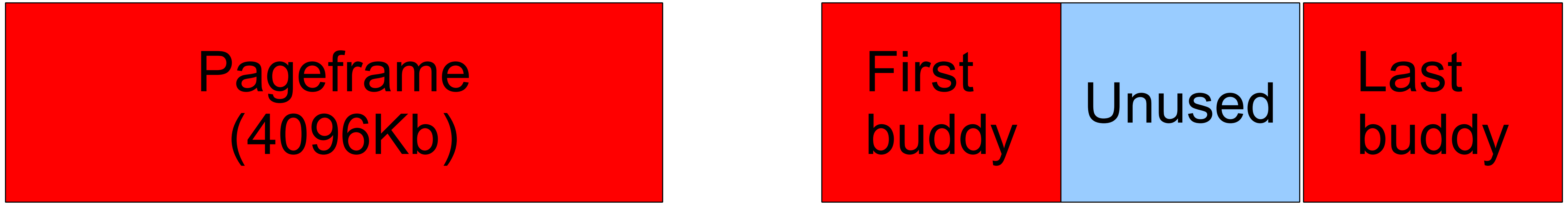
ZSWAP

ZBUD

Swap disk

ZBUD

- The allocator used by ZSWAP
- Store compressed pages
- Only allocate 0-order pages
- Pairs of zpages are “buddied”, one at the front of pageframe and one at the end
- No more than two zpages/buddies per pageframe
- Always search for the best fit buddy(the least wasted space)
- Page frames are LRU-linked

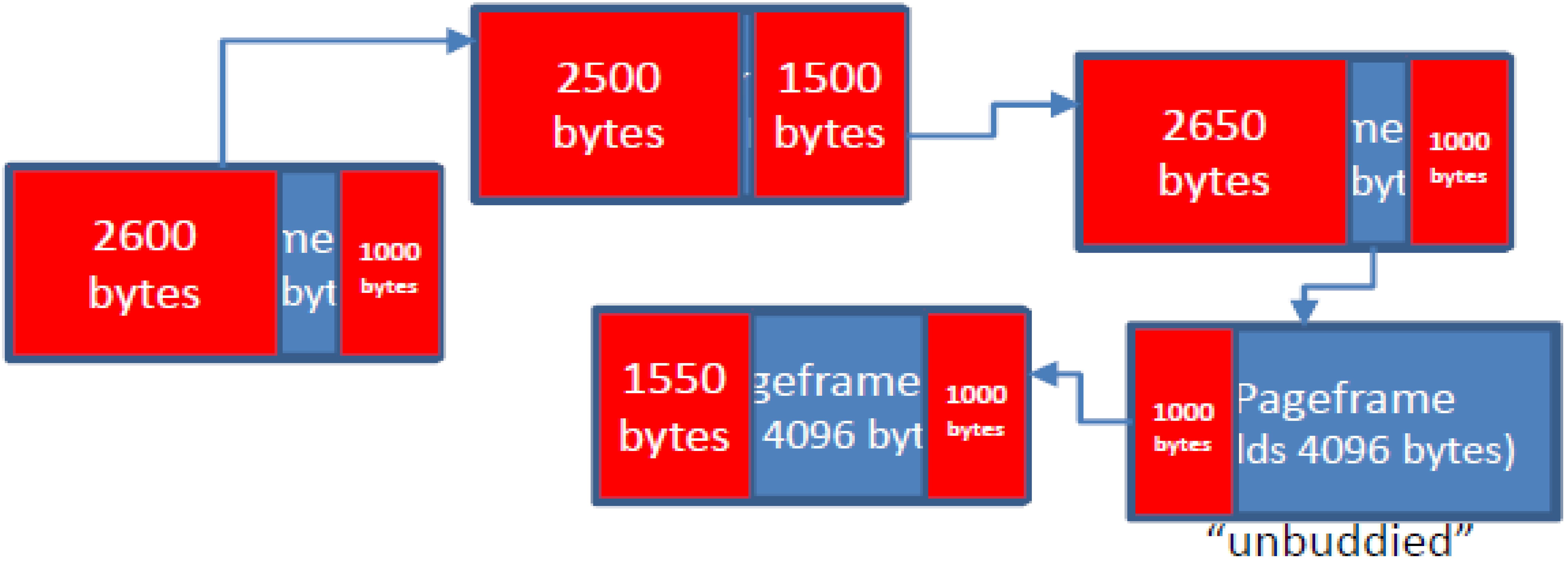


Pageframe (4096Kb)

First buddy

Unused

Last buddy



2500 bytes

1500 bytes

2600 bytes

ne byt

1000 bytes

2650 bytes

ne byt

1000 bytes

1550 bytes

geframe 4096 byt

1000 bytes

1000 bytes

Pageframe lds 4096 bytes)

"unbuddied"

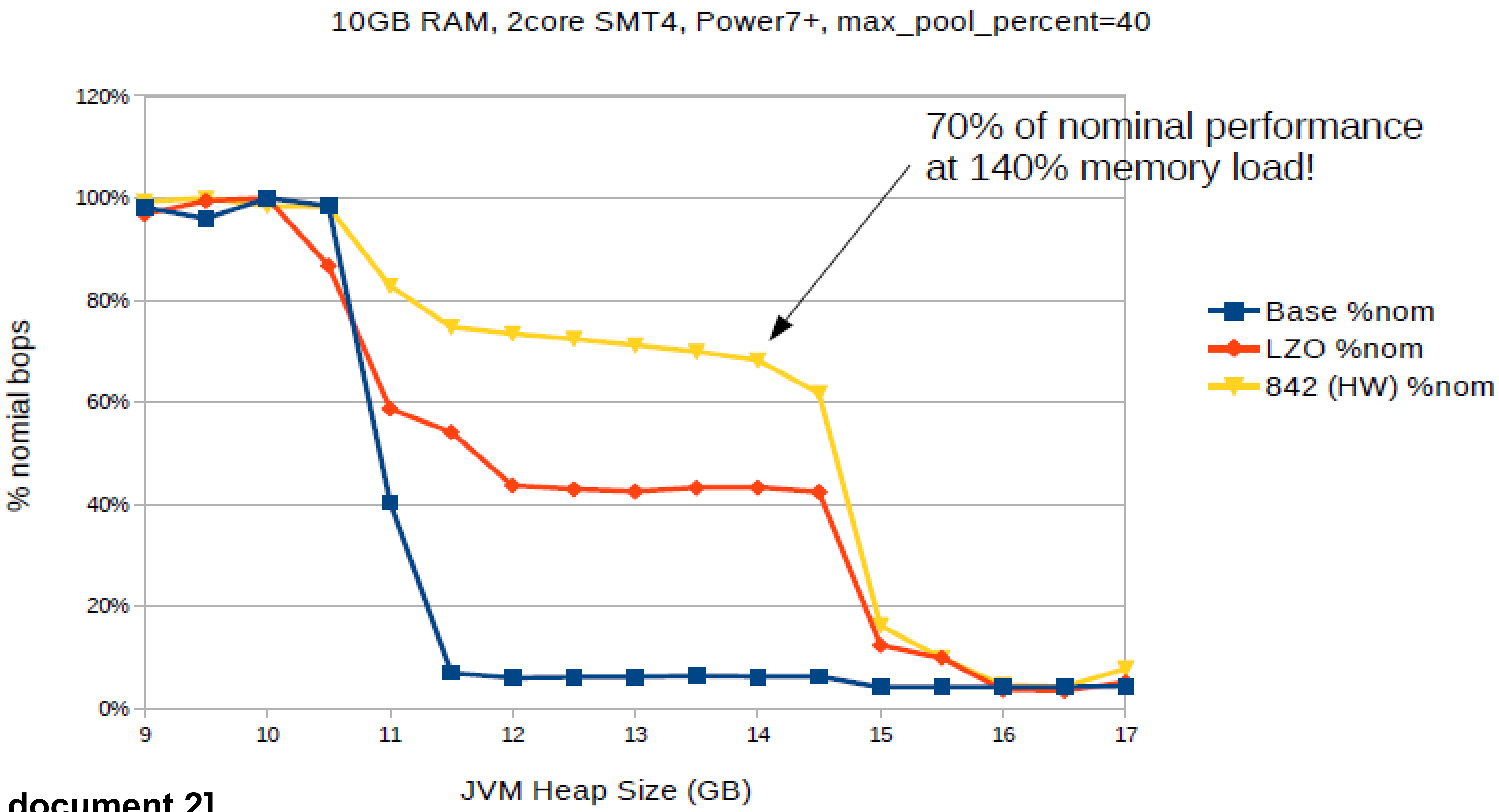
ZSWAP/ZBUD Status

- Get merged in v3.11
- mm/zswap.c
- mm/zbud.c

How to use ZSWAP

- Enable at boot time with a kernel parameter “zswap.enabled=1”
- Option parameters
  - zswap.compressor(lzo is default, can use hardware compressor like 842)
  - zswap.max_pool_percent
- Statistics
  - /sys/kernel/debug/zswap
  - /sys/kernel/debug/frontswap

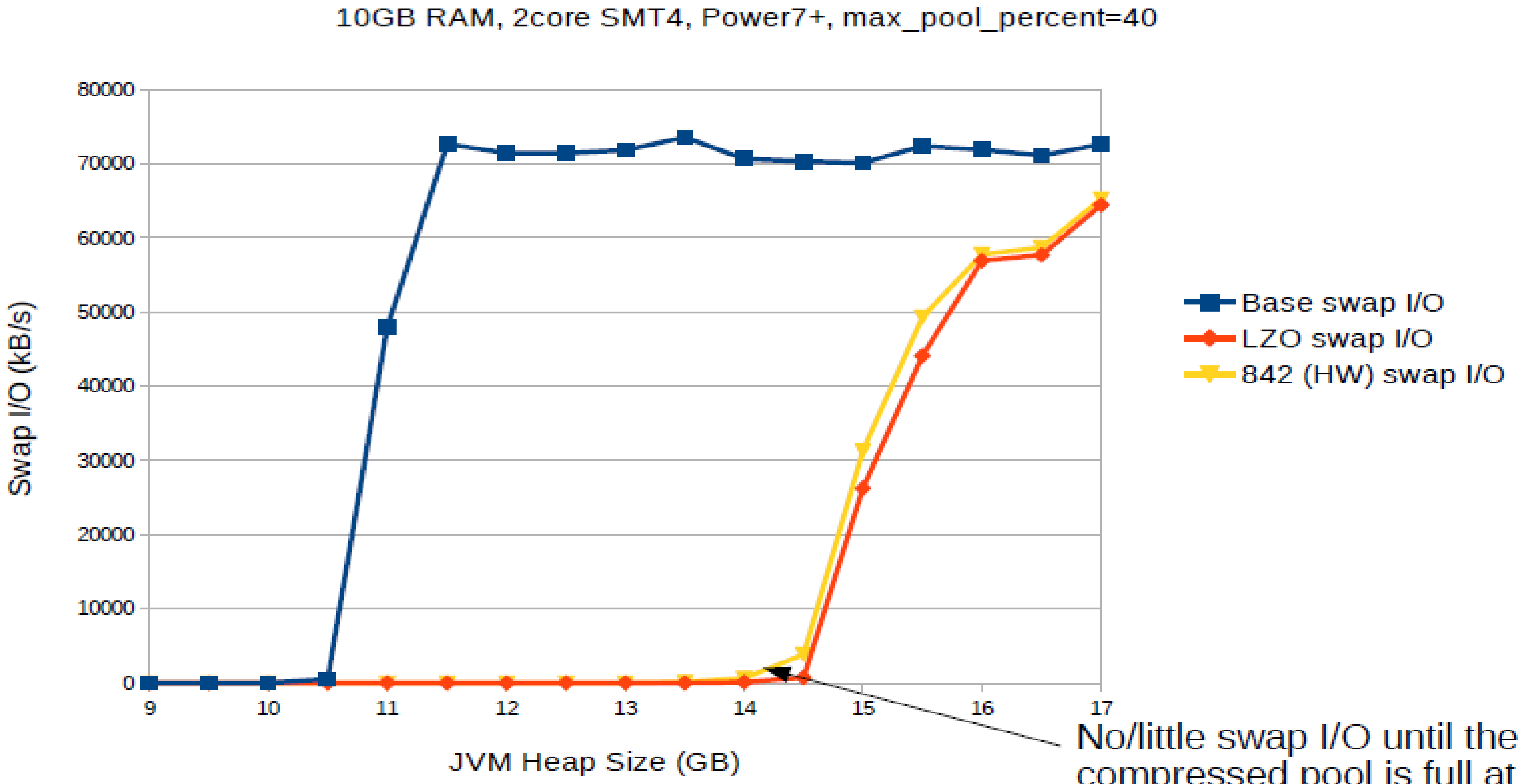
SPECjbb Performance



70% of nominal performance at 140% memory load!

* [Reference document 2]

Swap I/O



No/little swap I/O until the compressed pool is full at 140% memory load

* [Reference document 2]

ZSWAP/ZBUD pageframe reclaim

- When memory pool(store compressed pages) hit the limitation(default 20%)
- Select the pageframe at the tail of zbud LRU list
- Decompress the pageframe into two new allocated pages
- Insert decompressed pages(two) into swapcache

ZSWAP/ZBUD pageframe reclaim

- Write these two decompressed pages to real swap device
  - VM page reclaim will recognize these two pages as clean pages and free them directly
- Free the pageframe used by zswap/zbud

ZSWAP/ZBUD pageframe reclaim ---- Challenge

- Free one pageframe acquires temporarily allocating two new pages
- Possible solution
  - Teach the core VM about zbud pages in the reclaim page

ZRAM

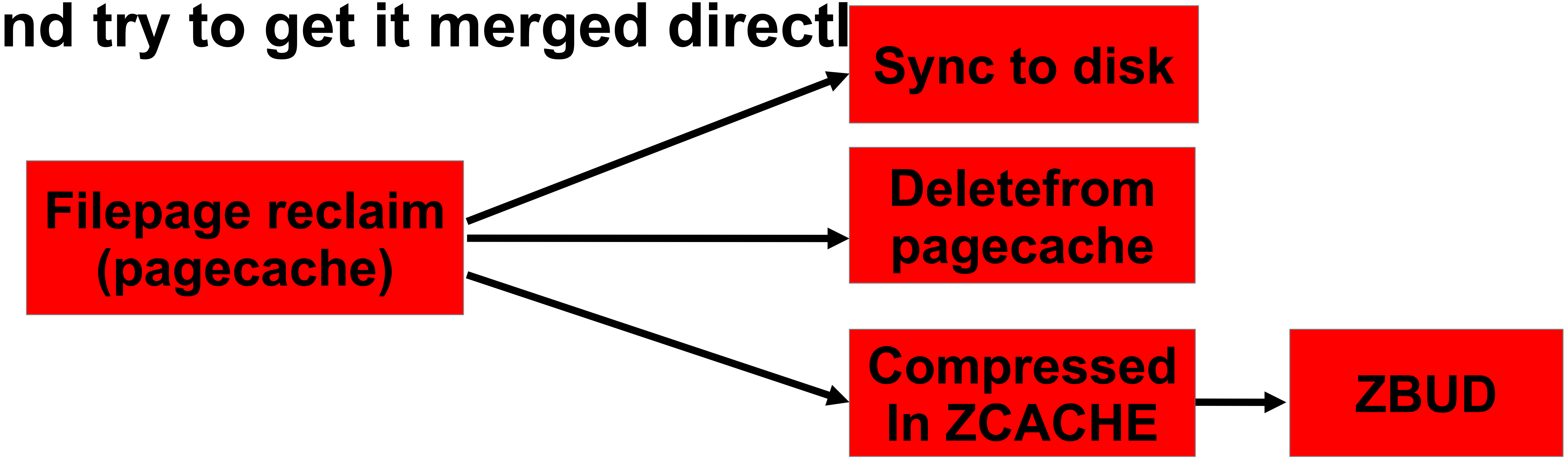
- Drivers/staging/zram/
- Act as compressed block ramdisk /dev/zram0
- Used as swap device or normal block device
- Use zsmalloc as the allocator which has high density but may with fragmentation issues (It may lead to unpredicable result)
- No pageframe reclaim
- Preferred by embedded system which may not have any real swap device

Merge ZRAM into ZSWAP or viceversa

- In theory but no agreement yet

ZCACHE

- A lot of things over the years in drivers/staing/zcache
- Dropped from staging recently
- My action: strip it down to only handle file page compression and try to get it merged directl



Help needed

- Real workloads performance

References

- 1. The zswap compressed swap cache
- 2. New Linux zswap compression functionality
- 3. Zcache: a compressed page cache
- 4. https://blogs.oracle.com/linuxkernel/

Q&A

QUESTIONS

ANSWERS