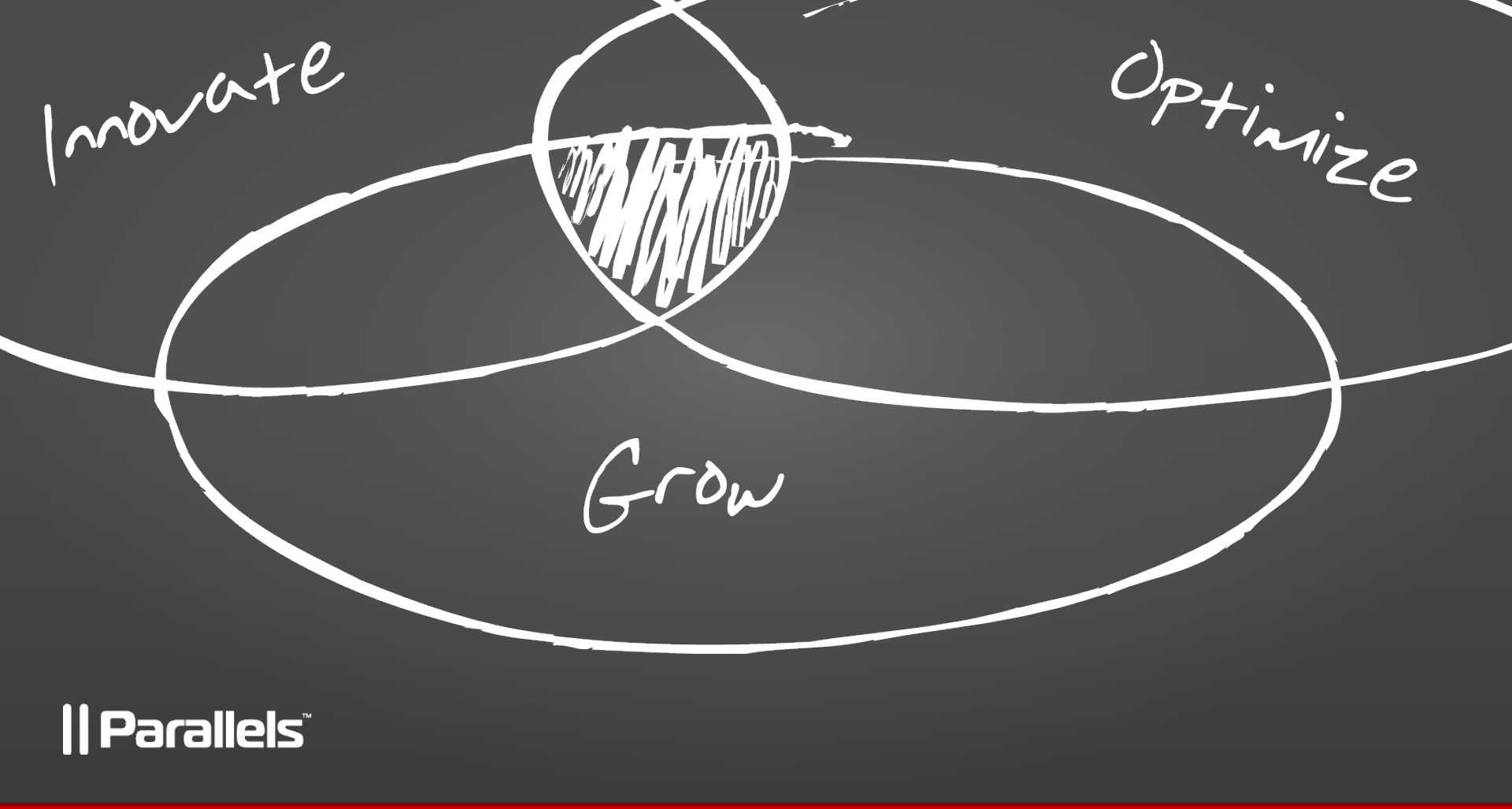
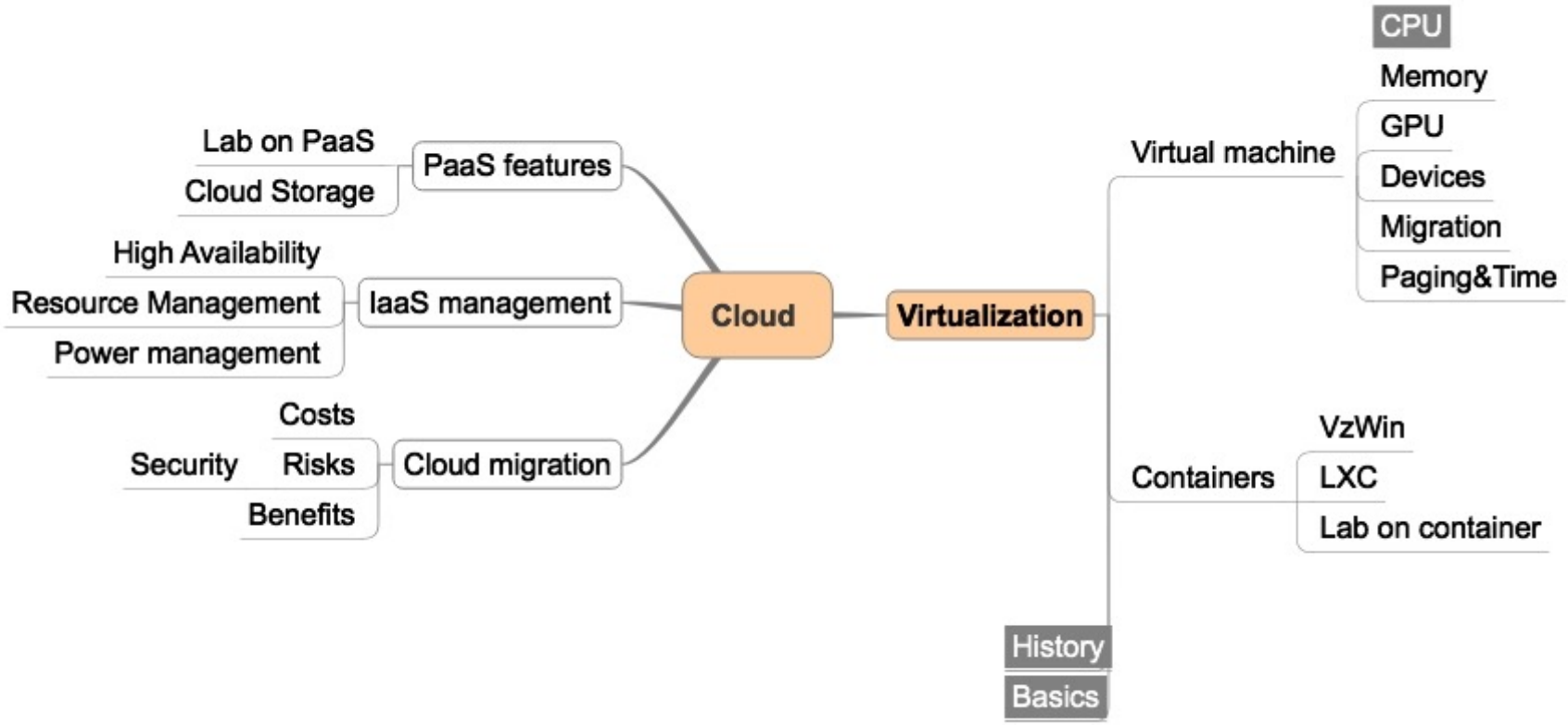


Parallels™

The total virtualization

Memory management

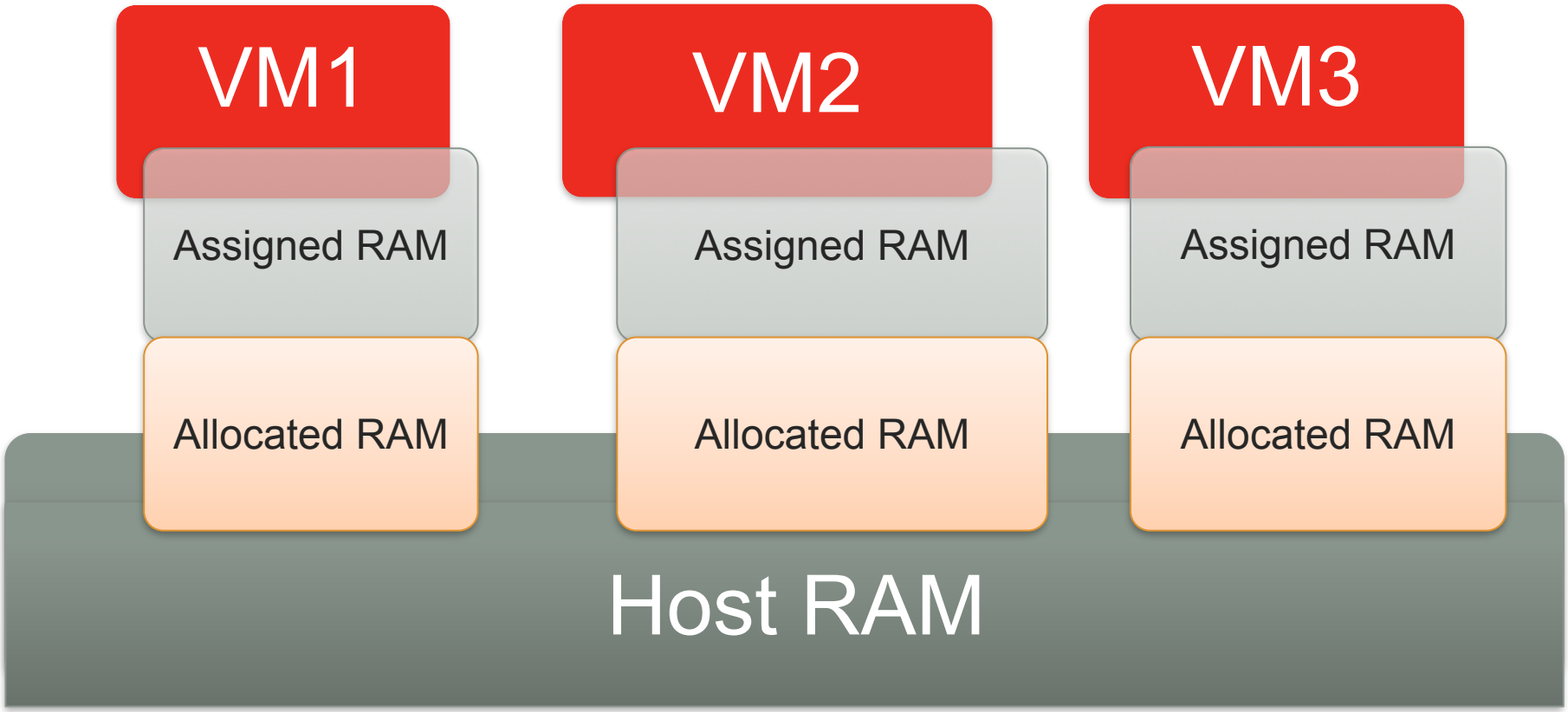
Course overview



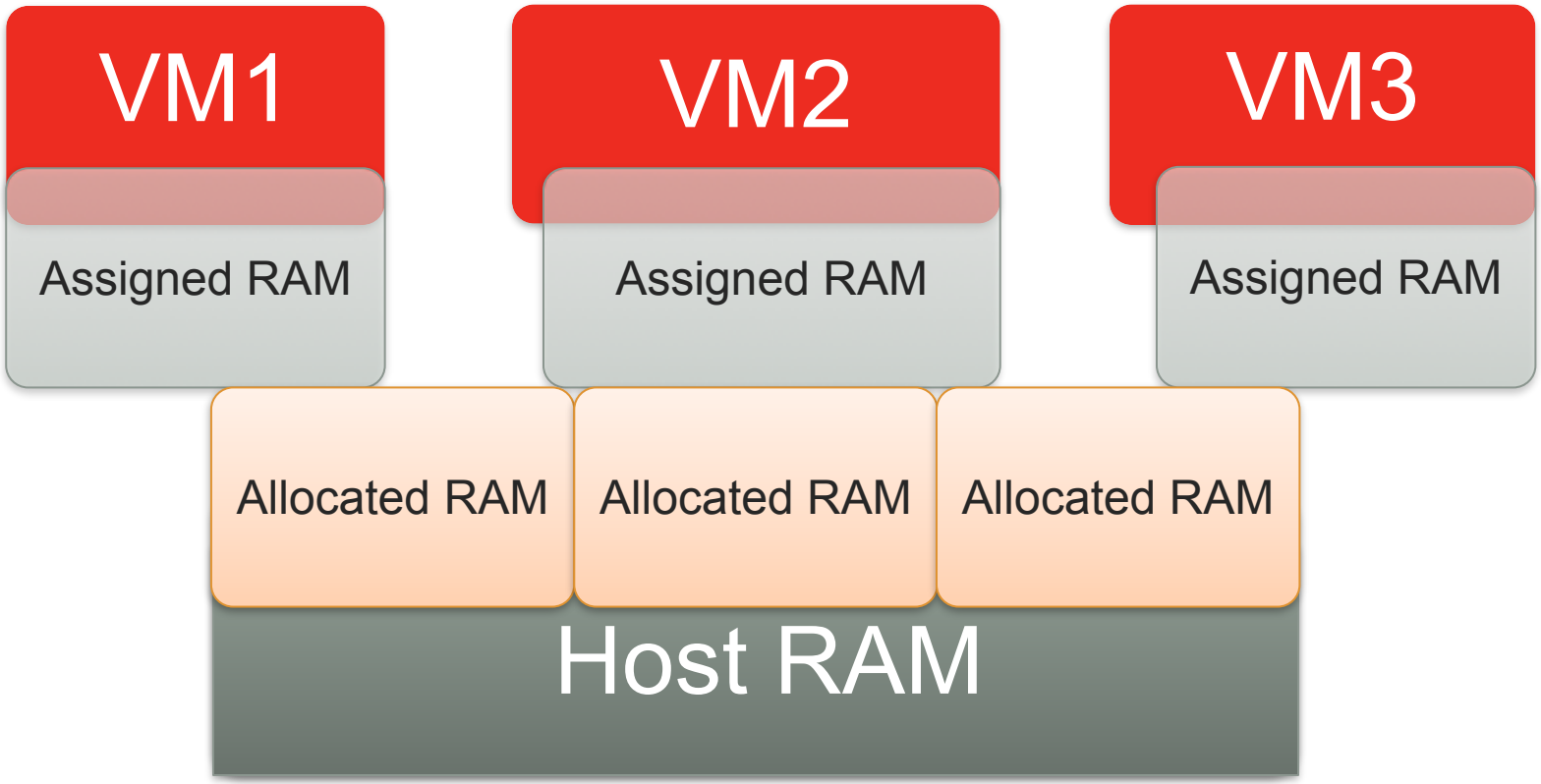
Memory management

- ✓ The task statement
- ✓ Technologies
  - ✓ VMM-swapping
  - ✓ Ballooning
  - ✓ Same page merging
  - ✓ Backing store choice
  - ✓ Memory compression
- ✓ The big picture

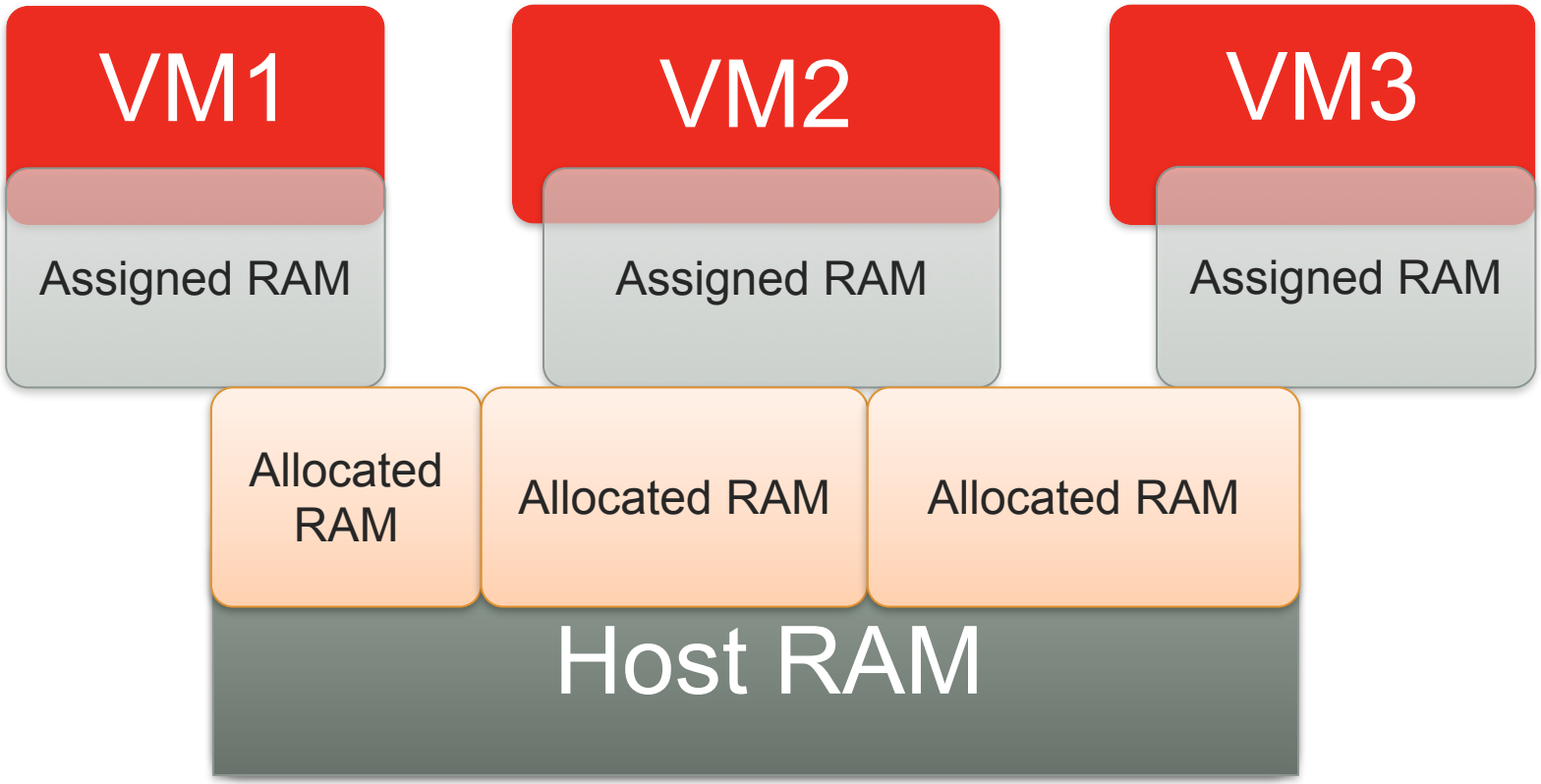
The task of memory management



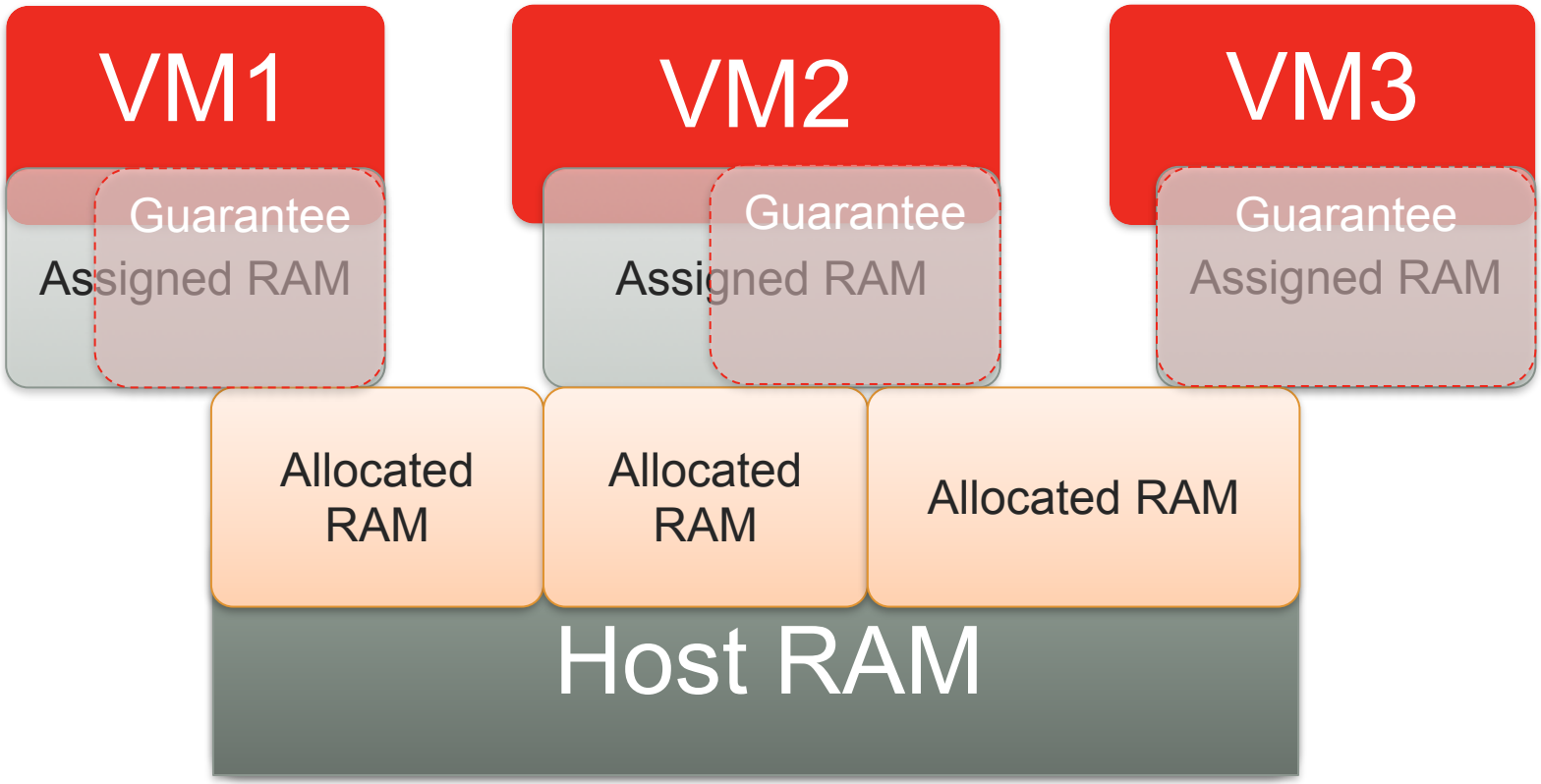
The task of memory management



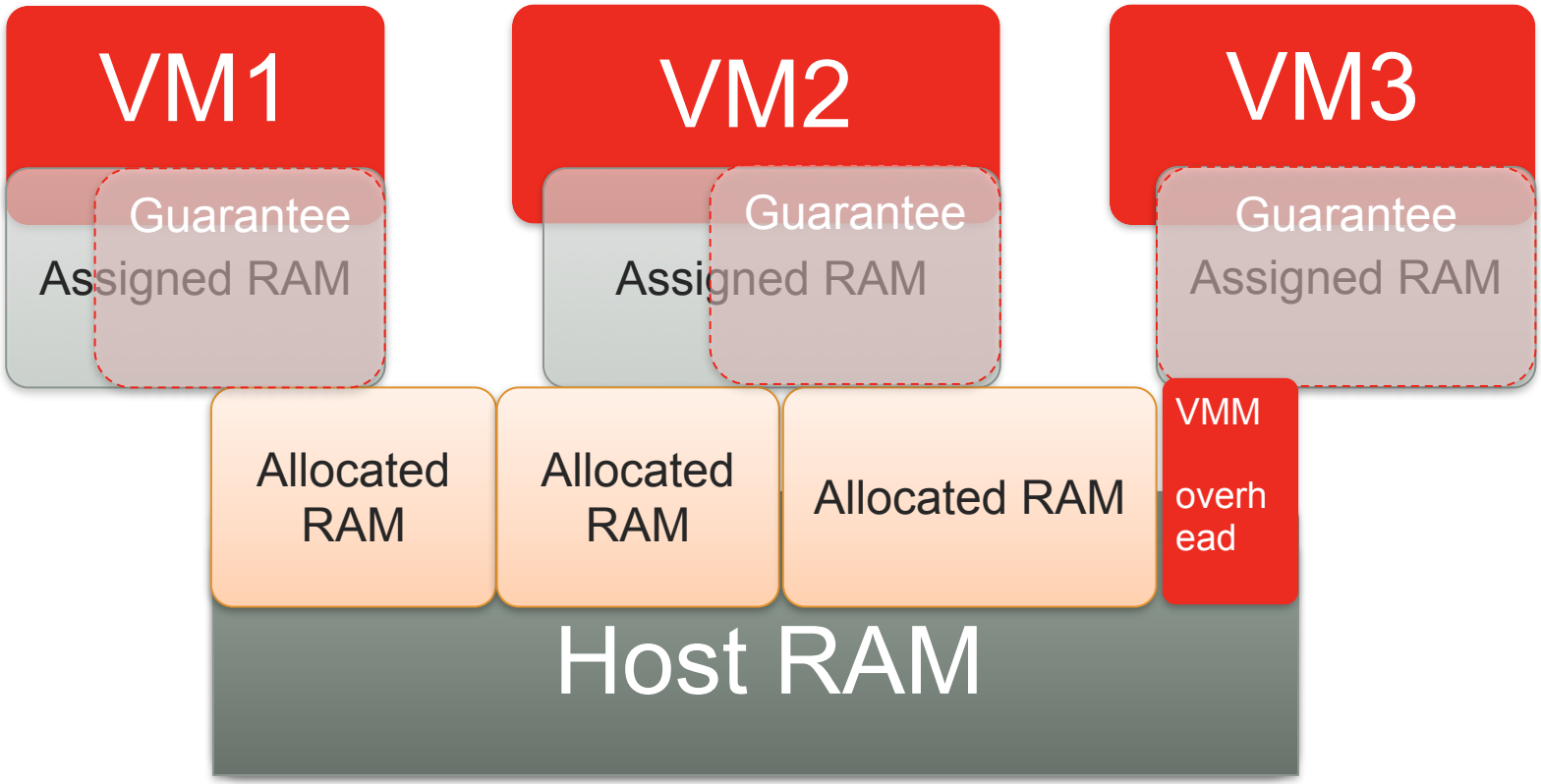
The task of memory management



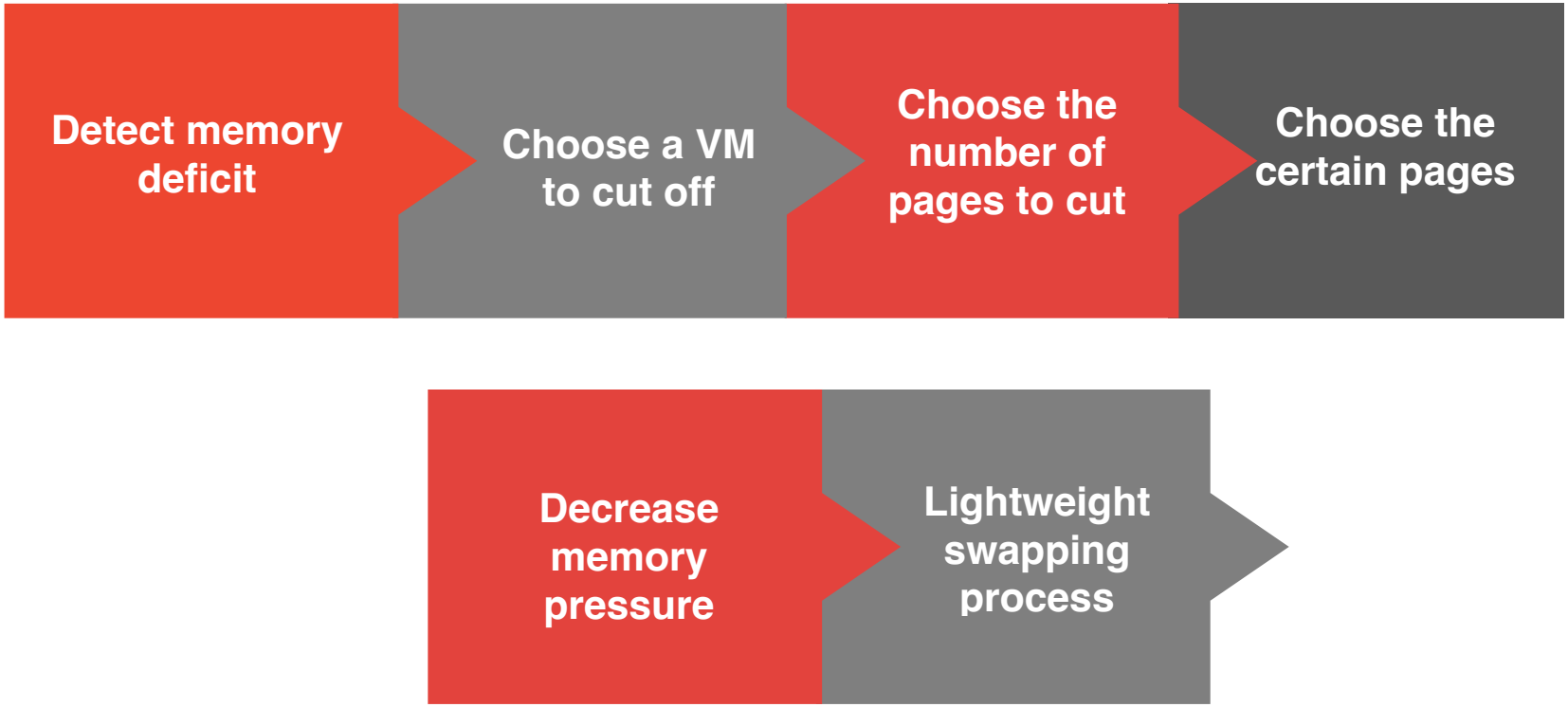
The task of memory management



The task of memory management



Subtasks of memory management



Managing host memory pressure

host memory pressure levels
Green, yellow and red zone/levels that makes swapping more aggressive

allowed guest consumption
VMs should consume below limit, otherwise they would be swapped

host OS memory pressure notifications
Subscribe for OOM-killer

continuous rebalancing
Looking for the harmony

Choosing VM to cut-off

max(current-guarantee)
Cut the VM that eats too much in comparison to original menu

share
Cut (WSS-guarantee)/(current-guarantee) proportionally to share

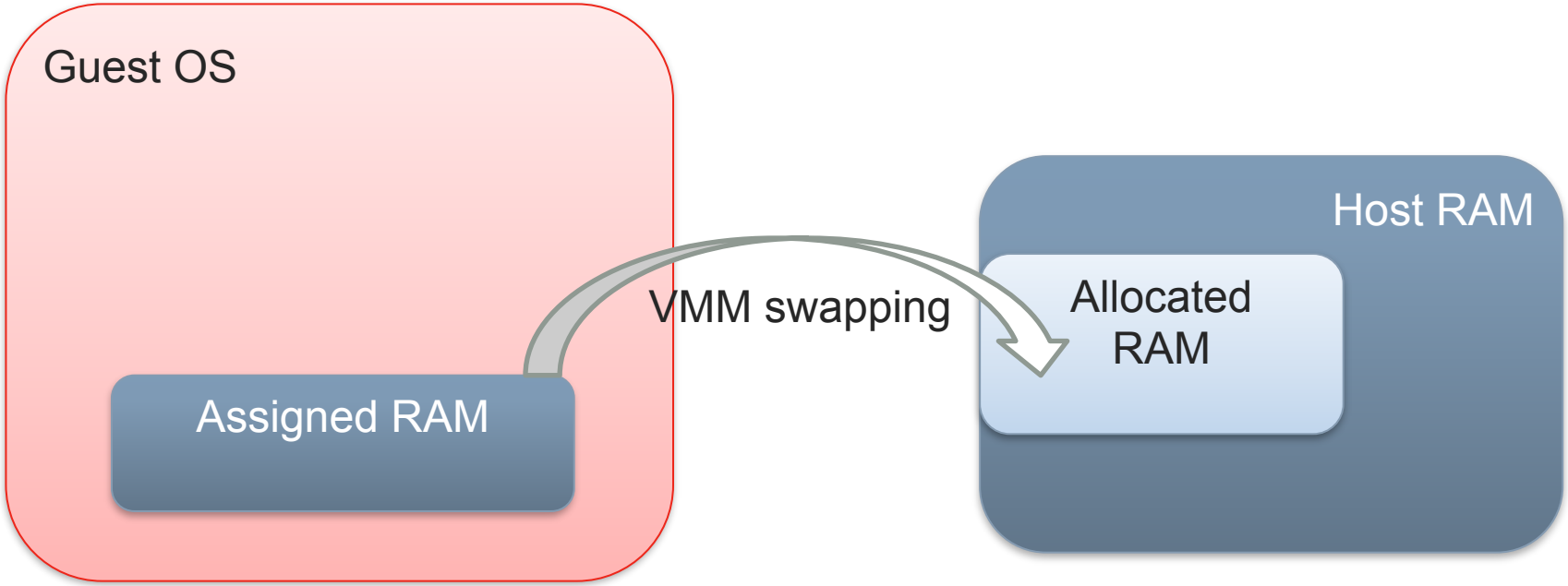
max(current-WSS)
WSS - what is WSS after all?

round-robin
Every VM would be cut, but let's do this in turn

idle tax
who doesn't work shall not eat ©

communism
Every time cut every VM

Memory management: the 1st approximation



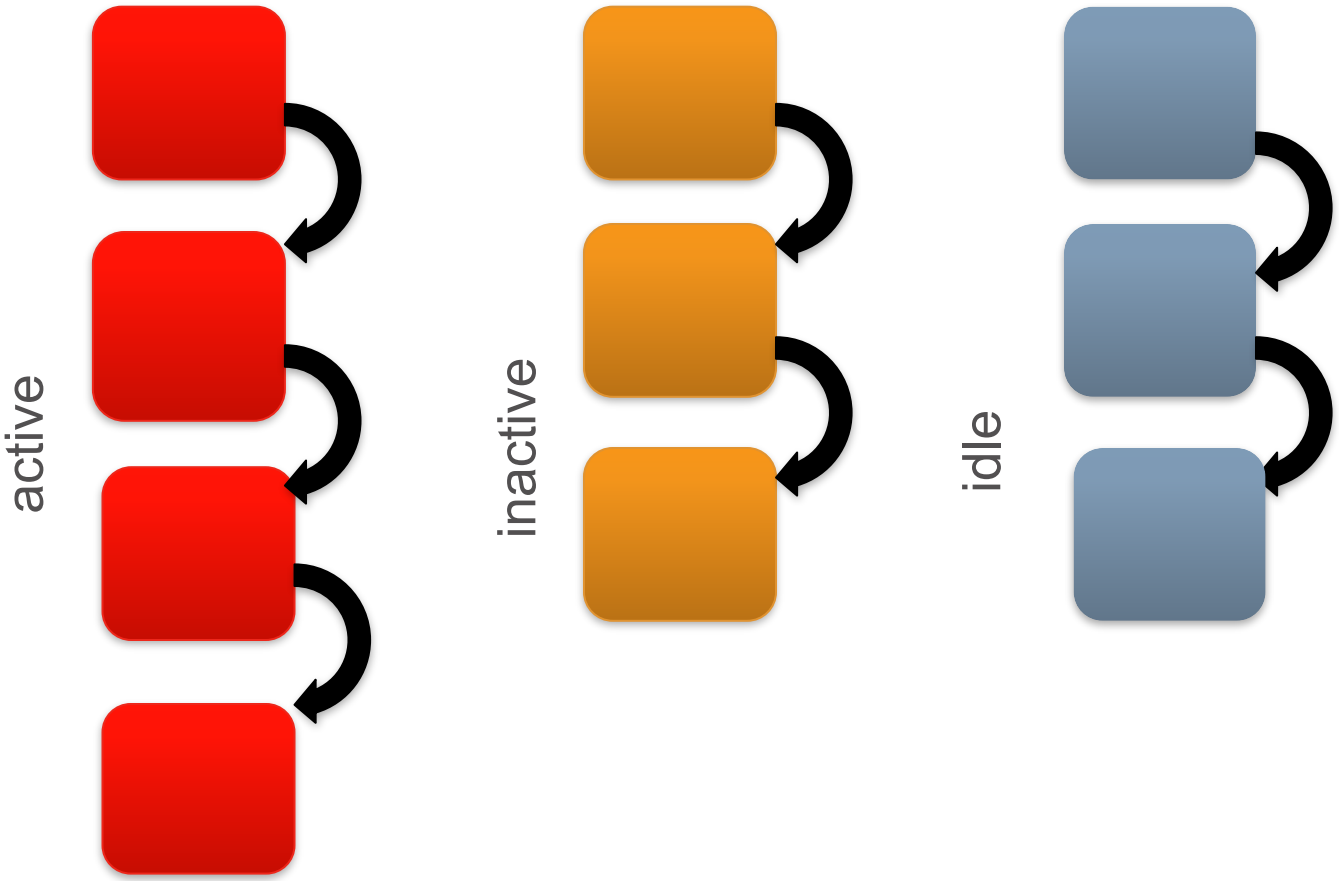
Replacement algorithms

- LRU (least recently used)
- …

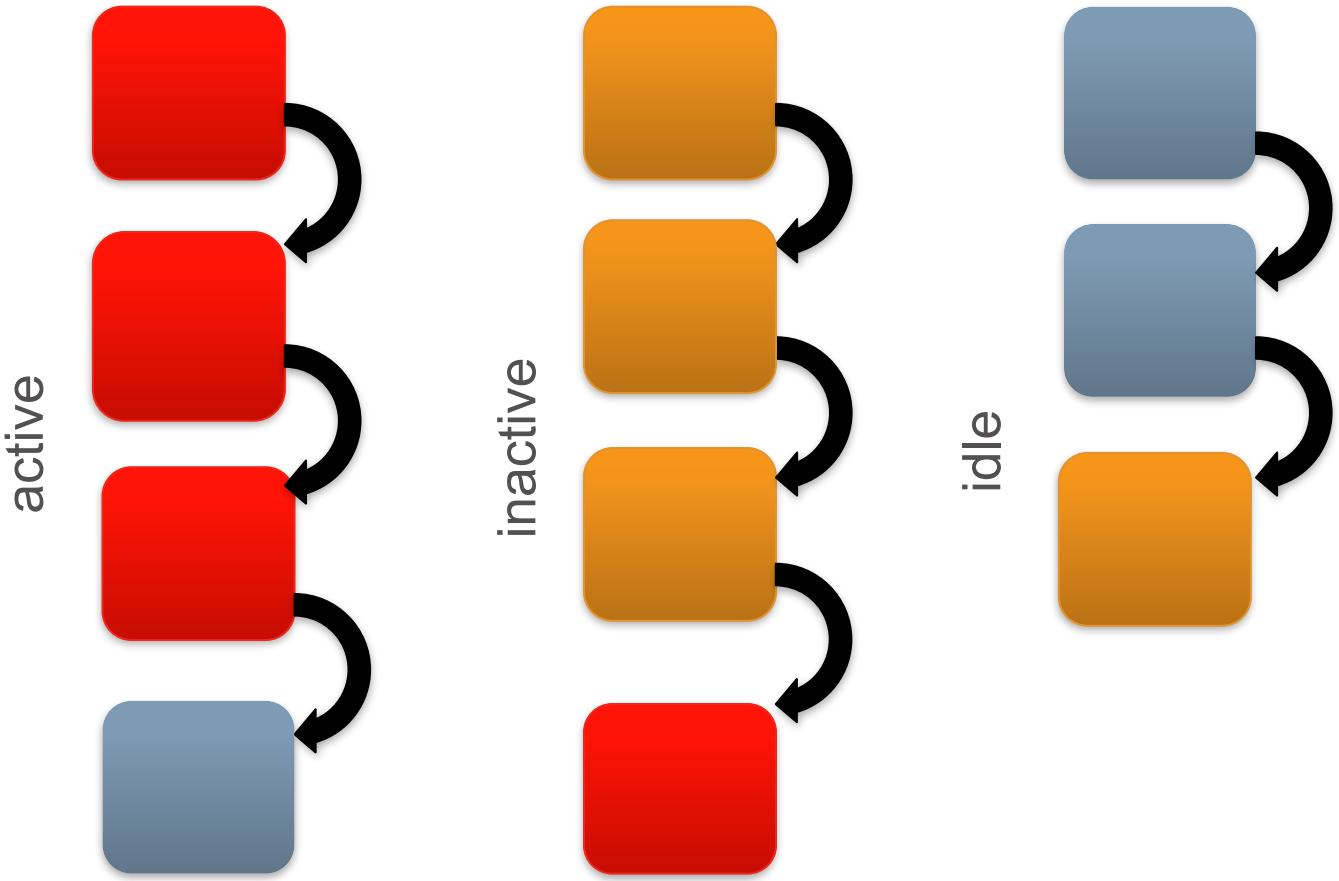
Replacement algorithms

- LRU (least recently used)
- FIFO (first in first out)
- NFU (not frequently used)
- Aging
- NRU (not recently used – A-/D- bits)
  - Second chance!
- WSClock
- Random (and random extrapolation)
- Frequency histogram
- Else?

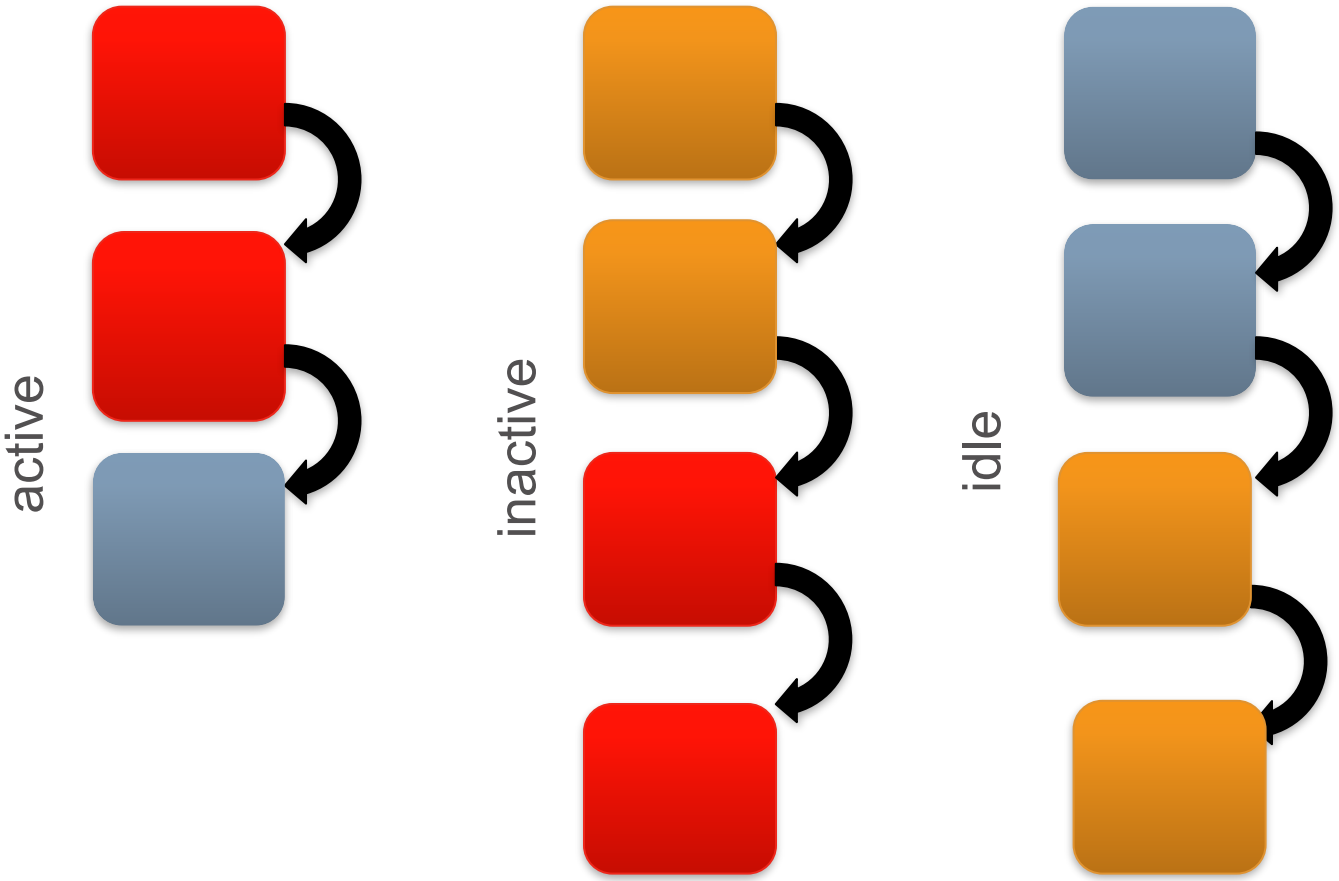
Replacement with a second chance



Replacement with a second chance



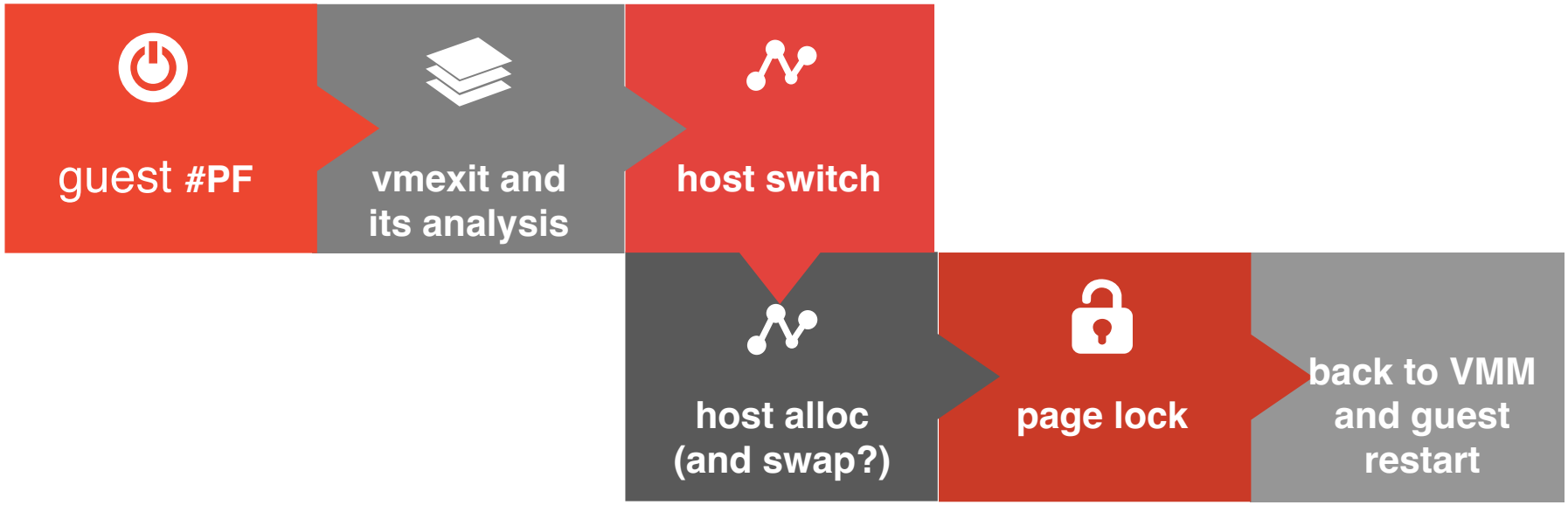
Replacement with a second chance



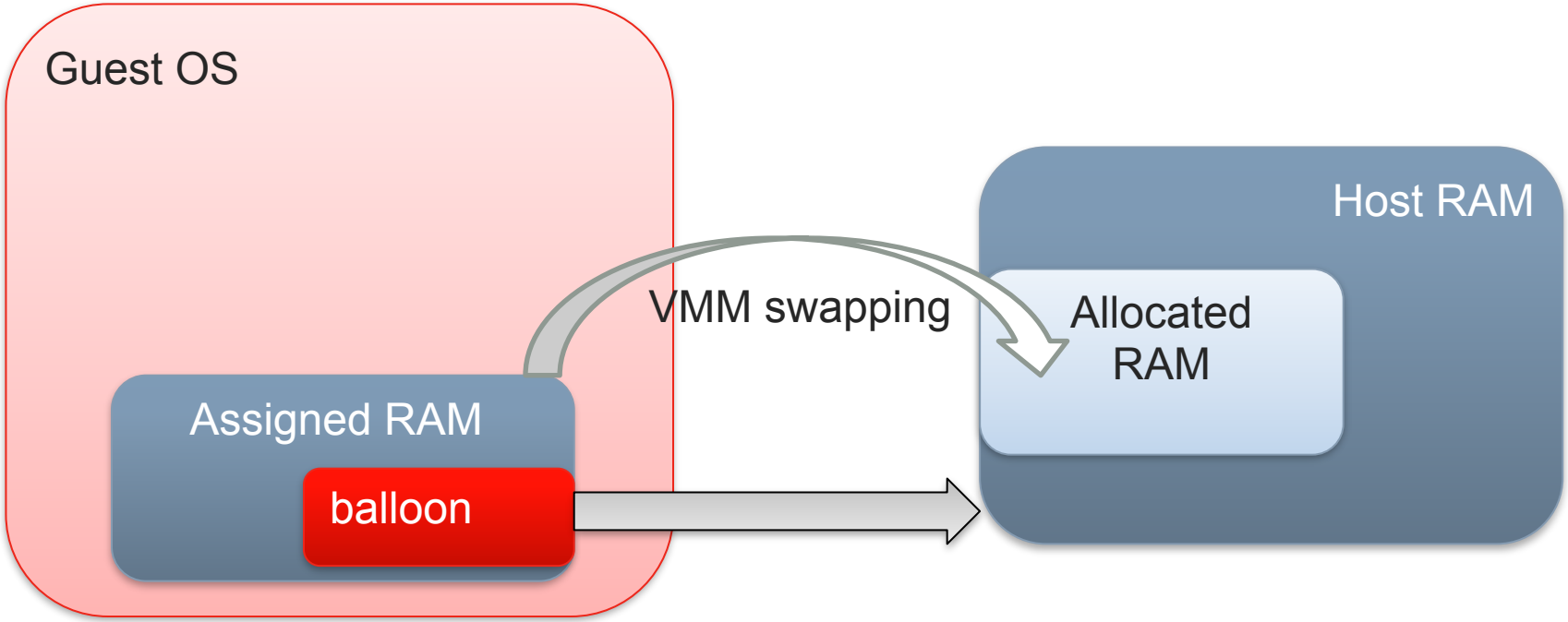
What's wrong with replacement?

- Semantic gap
- A-locality principle
- Large fine for the mistake (page miss makes VM suffer!)
- Insider's info

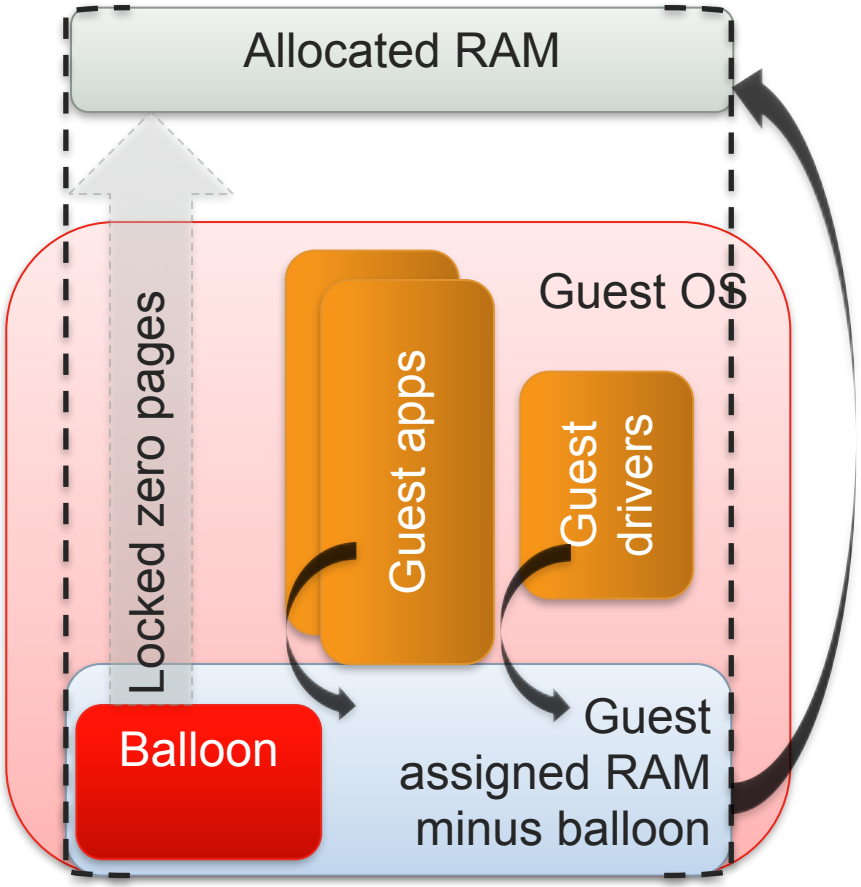
Page miss (VMware/Vbox/Parallels arch)



Memory management: the 2nd approximation



Ballooning



Balloon is an insider

Balloon's page won't be referenced by guest OS

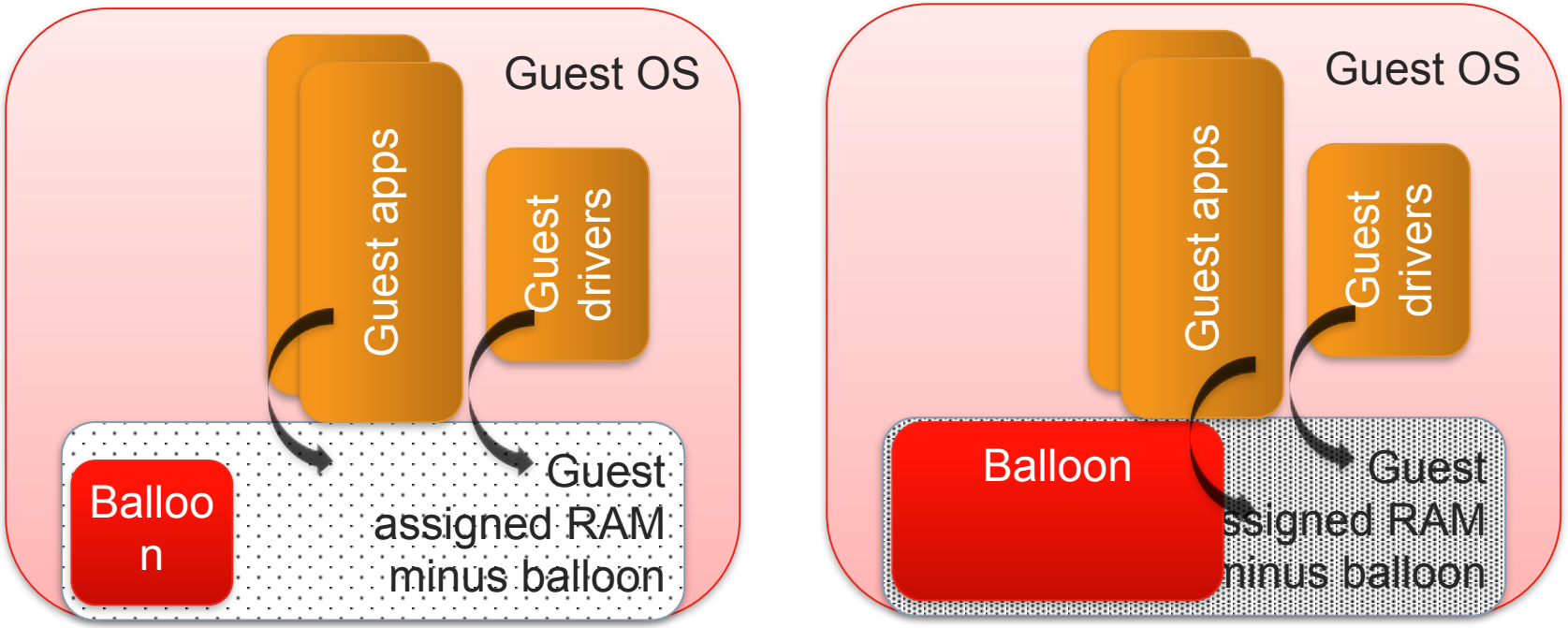
Balloon doesn't cause vmm swapping

Balloon pages have zero content (and shouldn't be stored)

Balloon is simple!

Ballooning

Balloon increases guest swapping, guest pressure

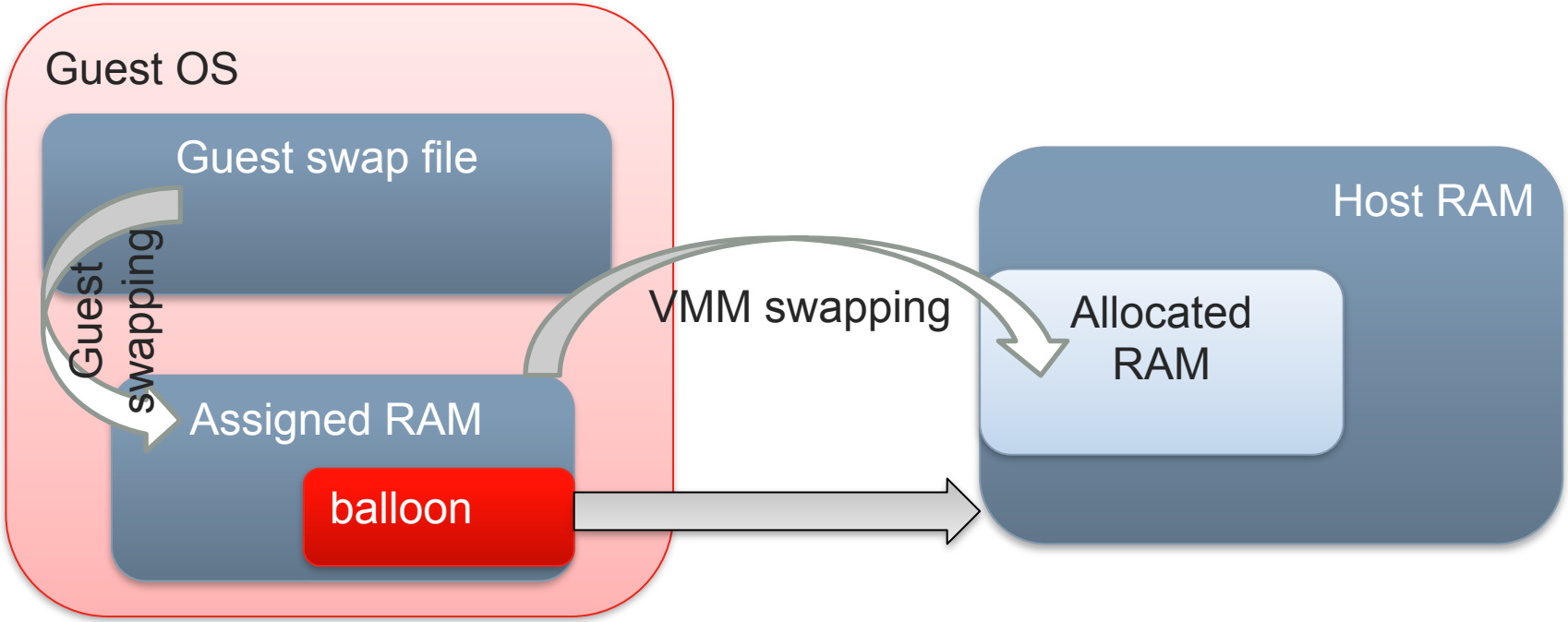


Balloon: No Silver Bullet

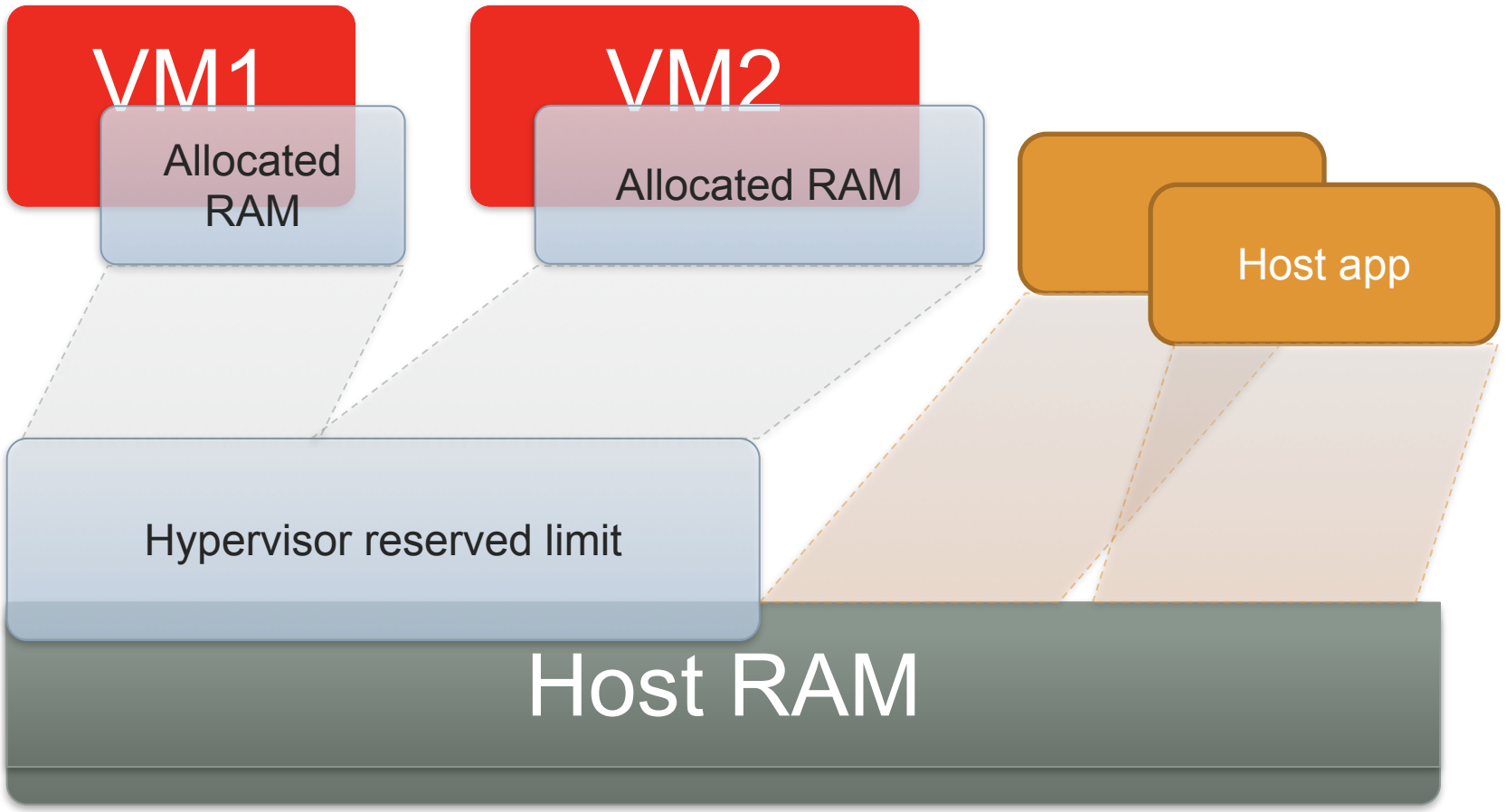
- Decrease resource usage with zero vmm swapping

- Guest swapping up to guest crashes (BSODs, OOM, etc)
- Need to re-implement the balloon for every guest system (if you need some modifications)
- No guarantees on balloon size
- When to deflate?
- The user could see the balloon

Memory management: the 3rd approximation



Memory quota



Memory quota

memory
installed(assigned) RAM

guarantee
minimum allocated memory

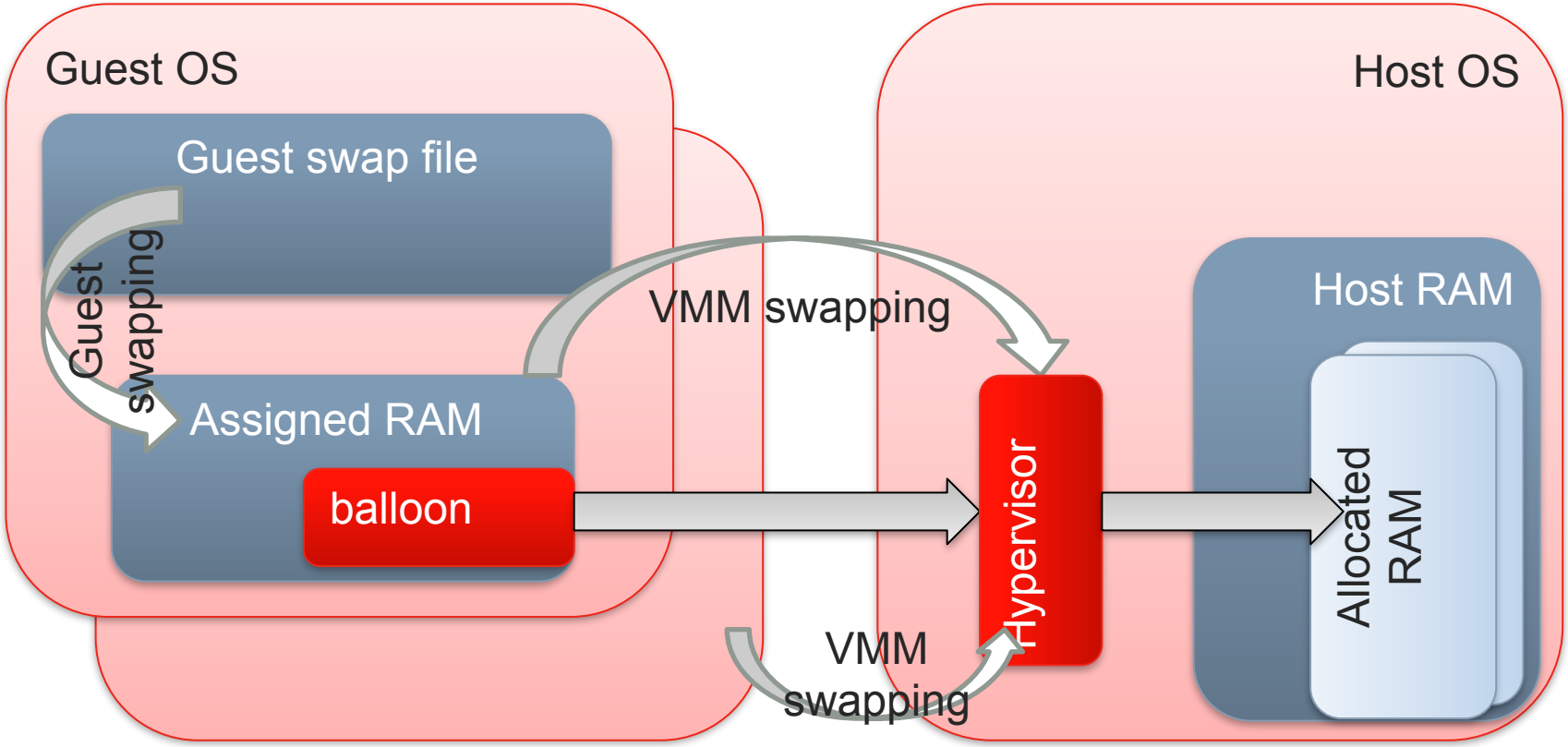
overhead?

limit
maximum allocated memory

share/priority
Which VM to cut off first/most?

maxmem
max allowed «hotplugged» memory

Memory management: the 4th approximation



How to reduce the overall memory usage

Transparent page sharing

Deduplication
(aka same page merging, KSM; aka THP)

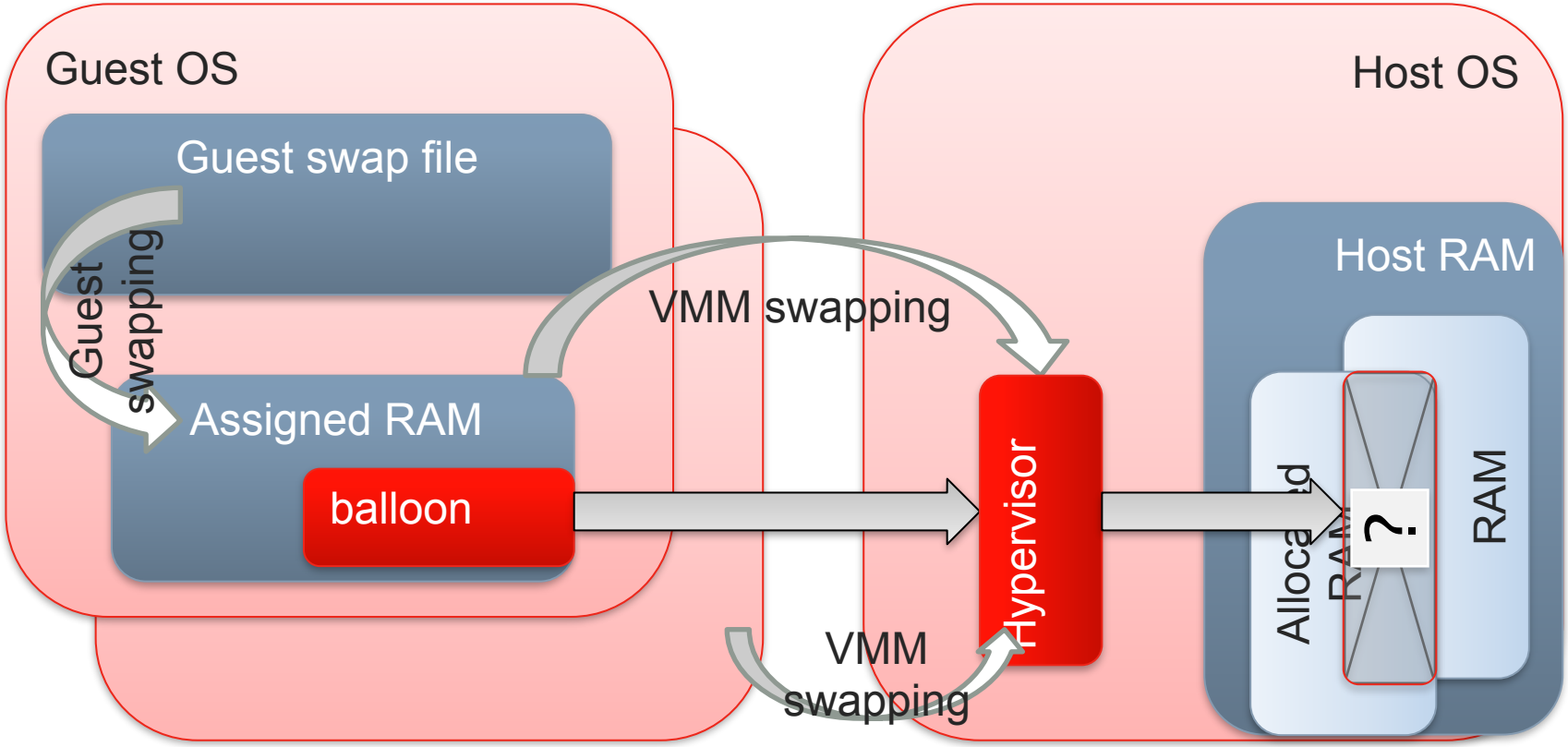
- Hash for each page (else $o(n^2)$ to compare)
- Search for equals (hash + cmp)
- Multiple virtual pages to point at one physical
- COW (copy on write)

Deduplication

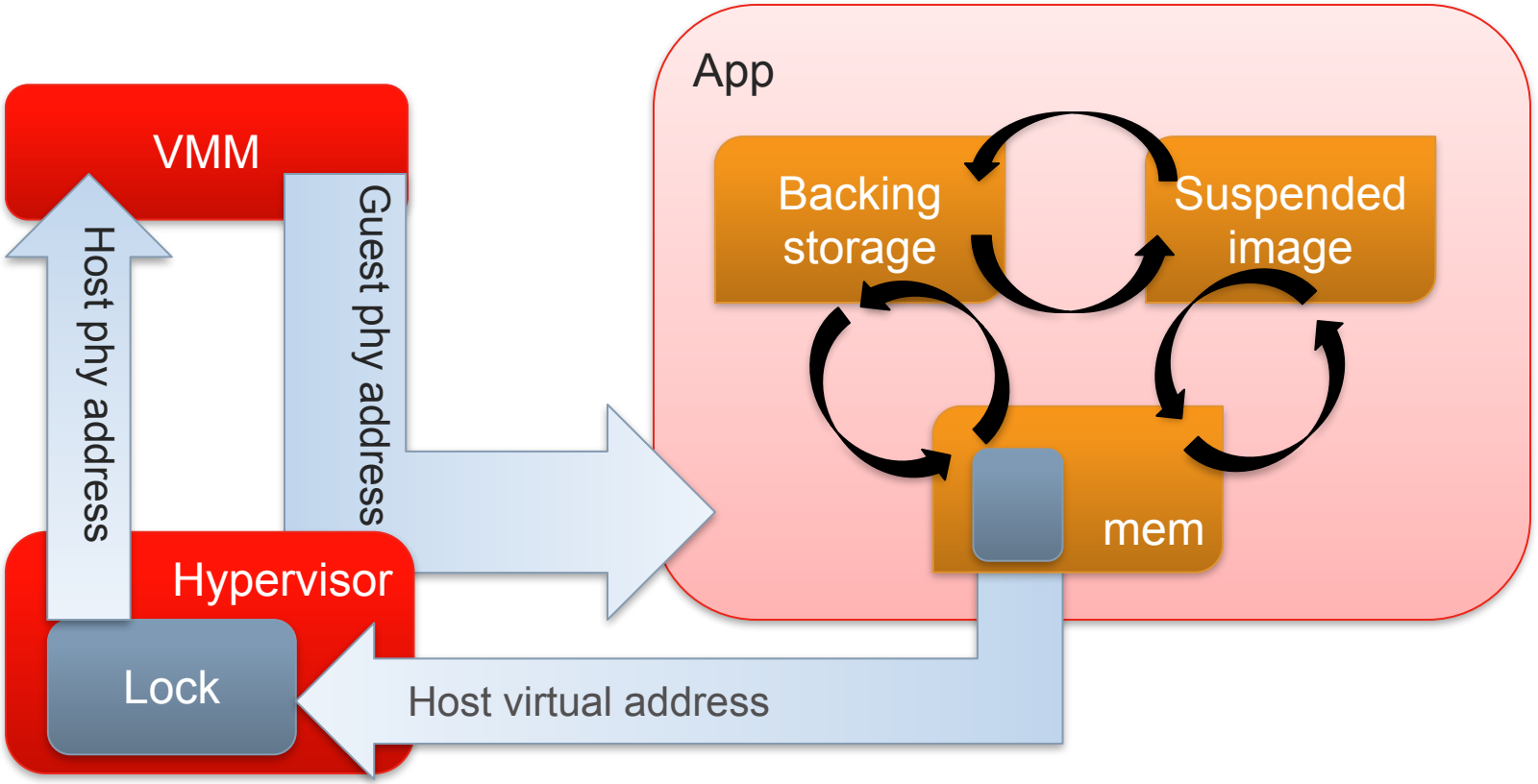
- Great for tests
- Enabled by default

- When to turn on (it introduces the guaranteed overhead and it doesn't guarantee any memory gain)
- When to turn off
- How to store hashes
- When to invalidate

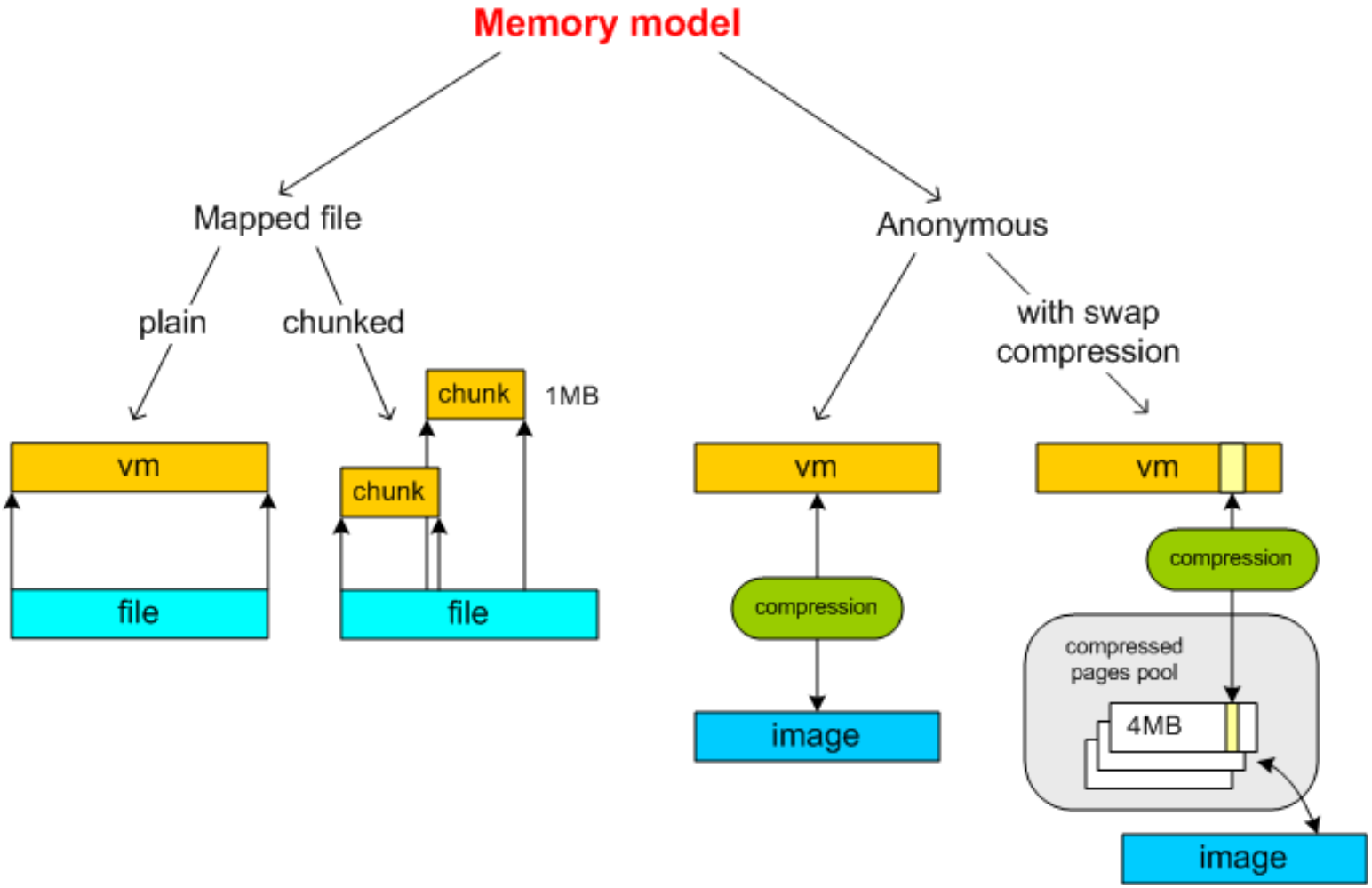
Memory management: the 5th approximation



Backing store model



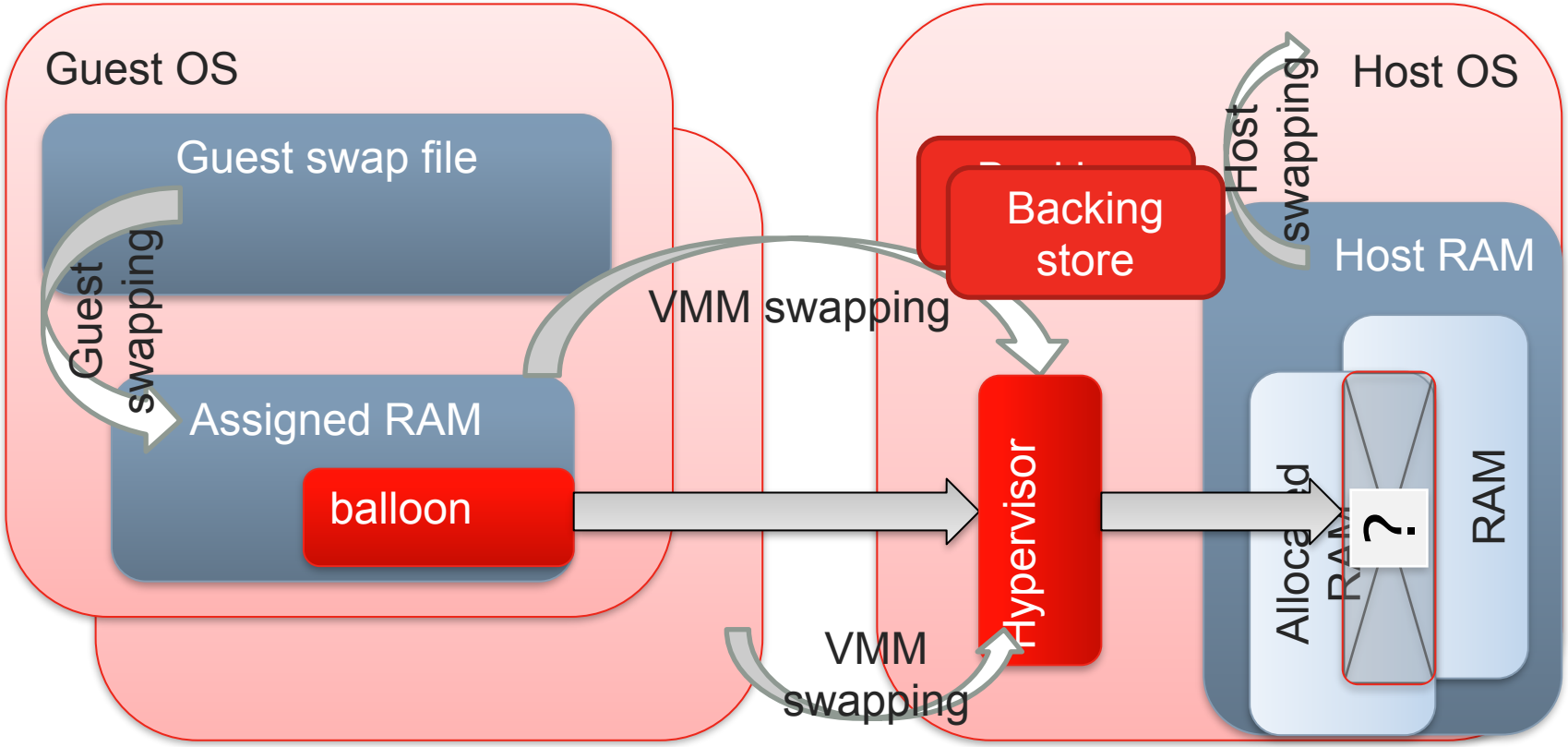
Backing store model



Backing store

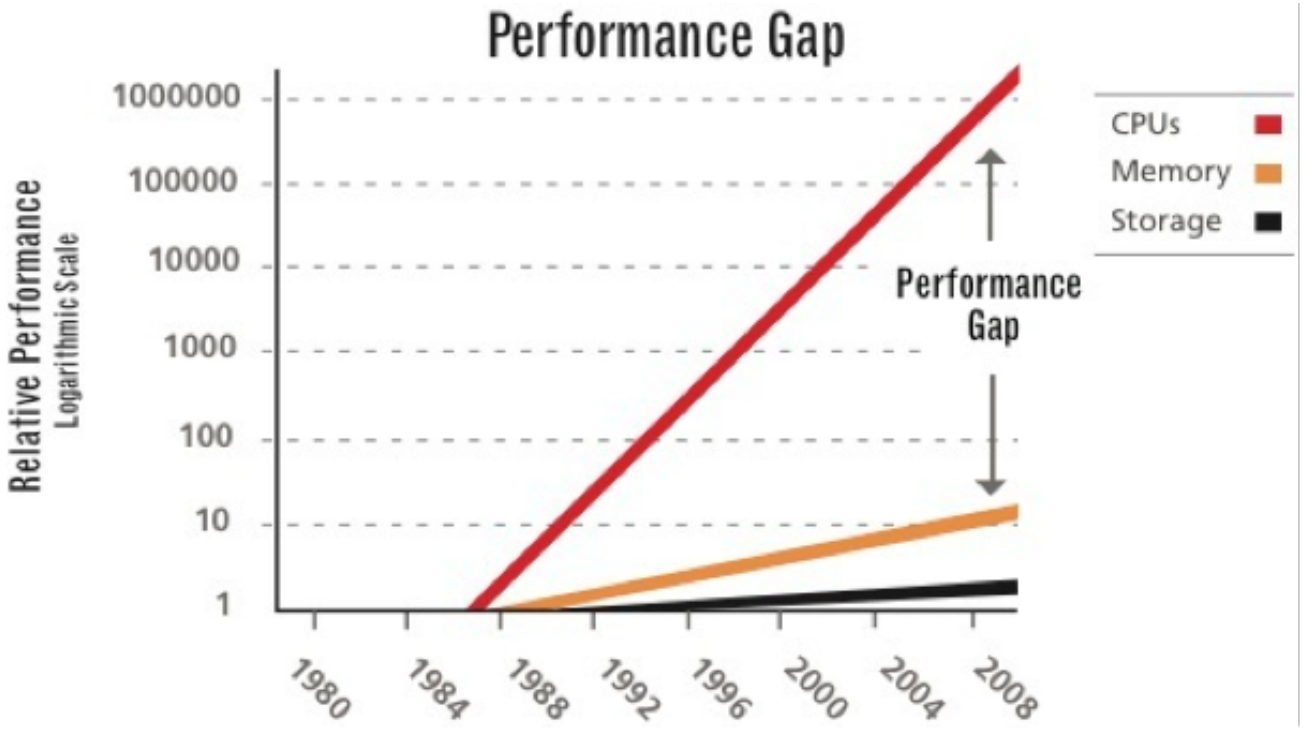
- Stores the content of swapped pages
- Supports suspend/resume/shapshot
- Supports migrate
- Allows access from both kernel space and user space

Memory management: the 6th approximation



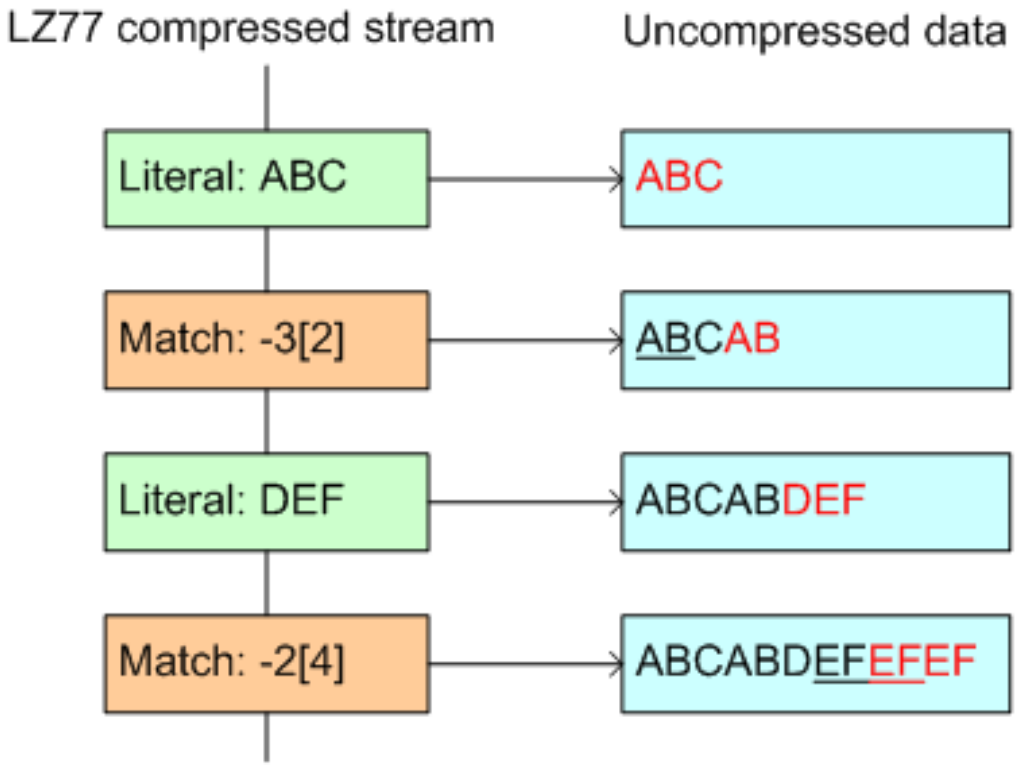
Compression

Compression: performance gap



© http://www.fusionio.com

Compression

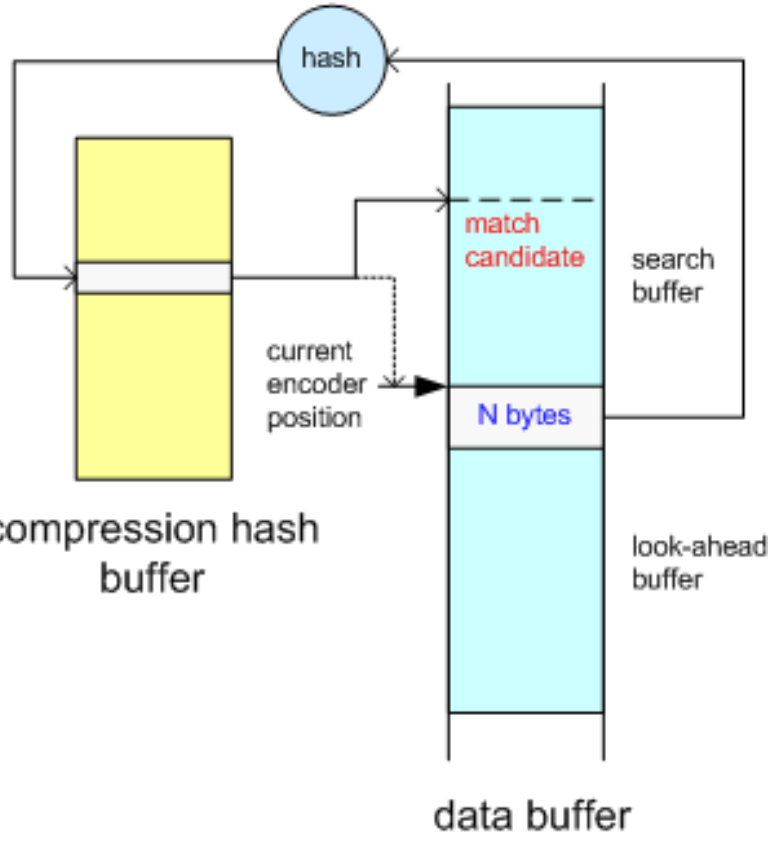


Jacob Ziv and Abraham Lempel; A Universal Algorithm for Sequential Data Compression, IEEE Transactions on Information Theory, 23(3), May 1977

Compression

LZRW = Lempel-Ziv Ross Williams

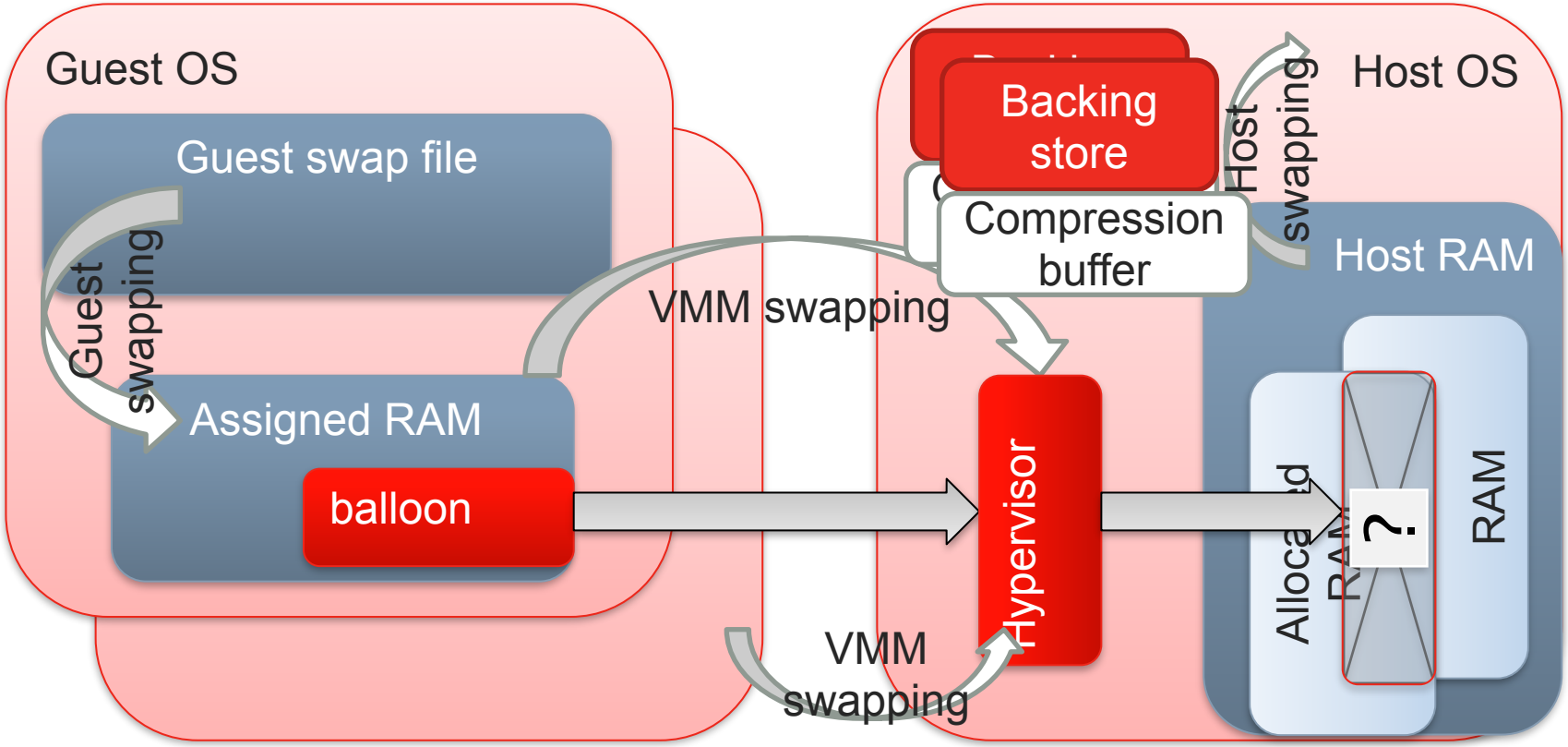
Williams, R.N., "An Extremely Fast Ziv-Lempel Data Compression Algorithm", Data Compression Conference 1991 (DCC'91)



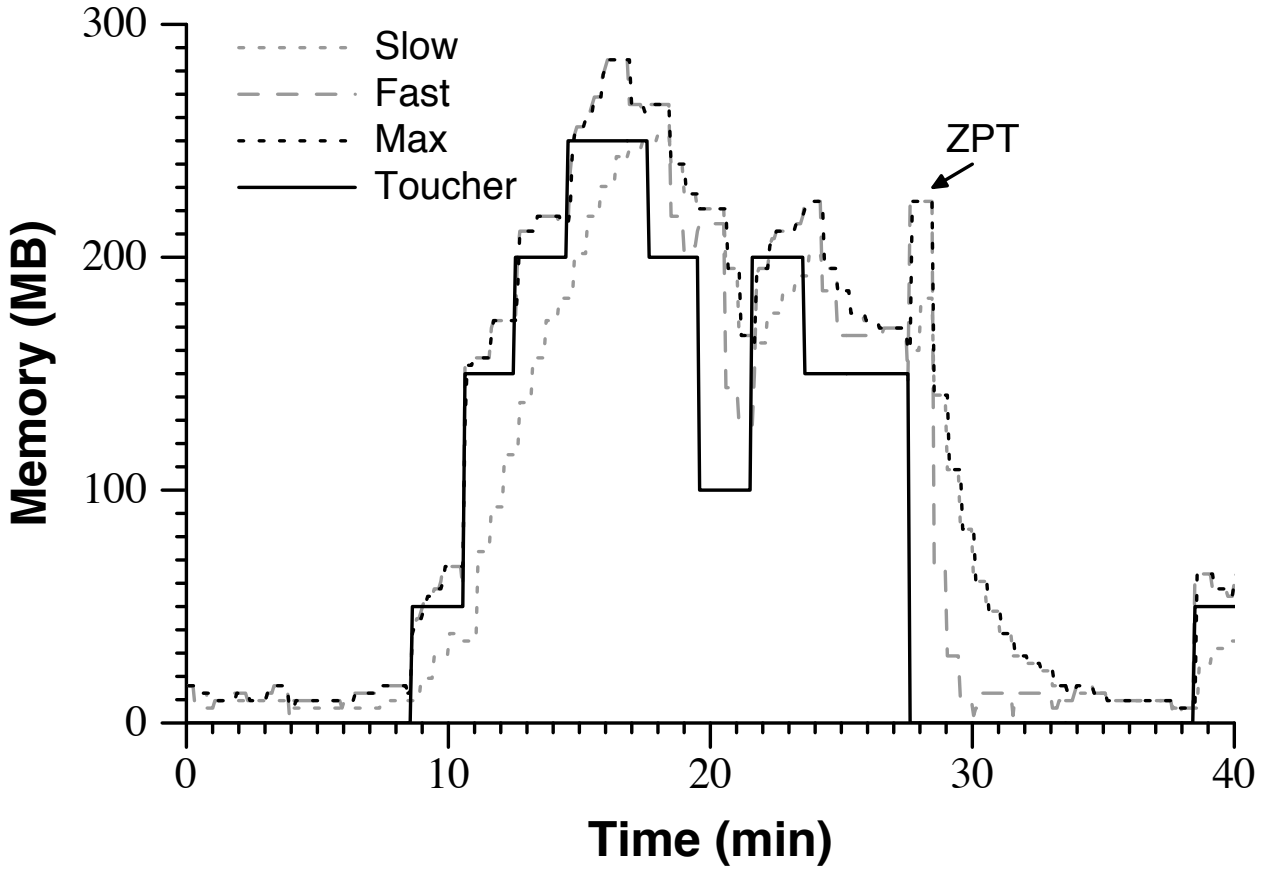
LZRW1: N=3, literals are marked by bitmap
PD6 ~200MB/sec compress/uncompress

LZRW4: N=4, literals are encoded by single byte tags
~250MB/sec compress
PD7 ~450MB/sec uncompress

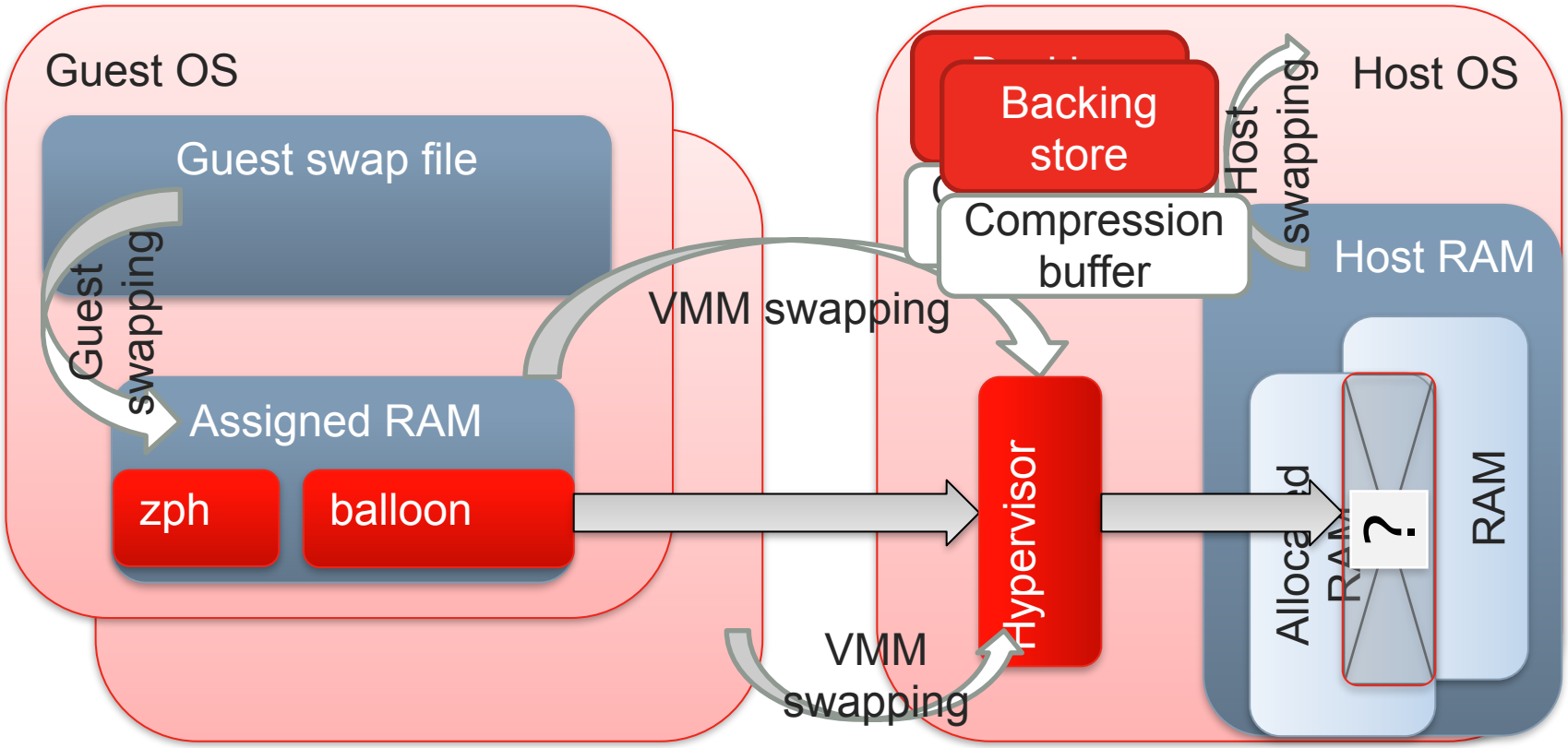
Memory virtualization: the 7th approximation



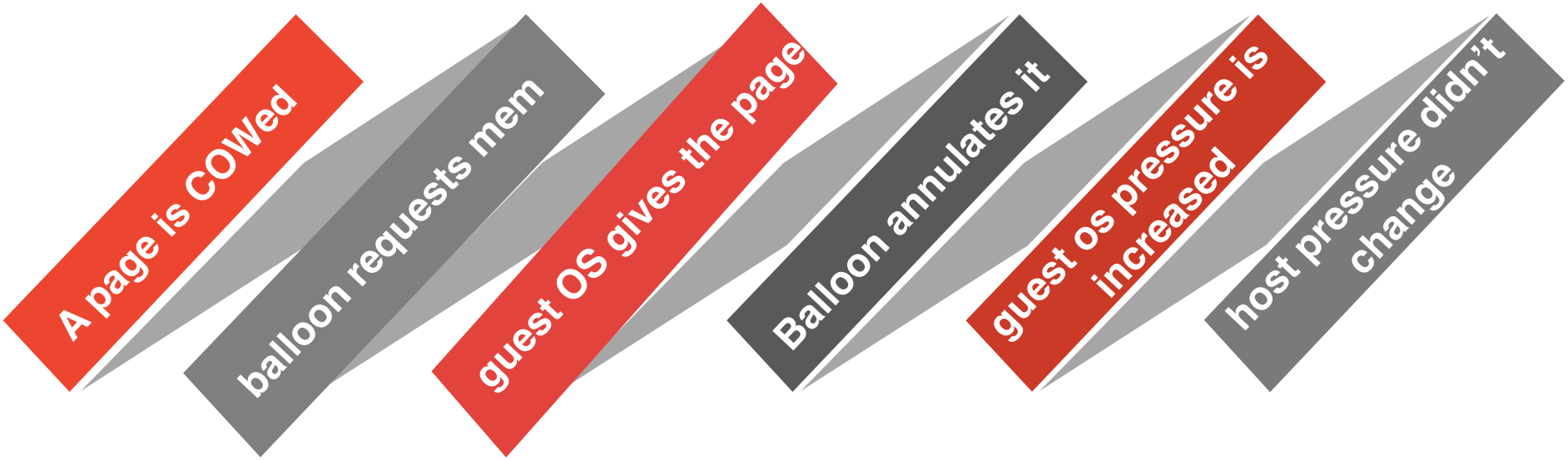
Zero Page Hack



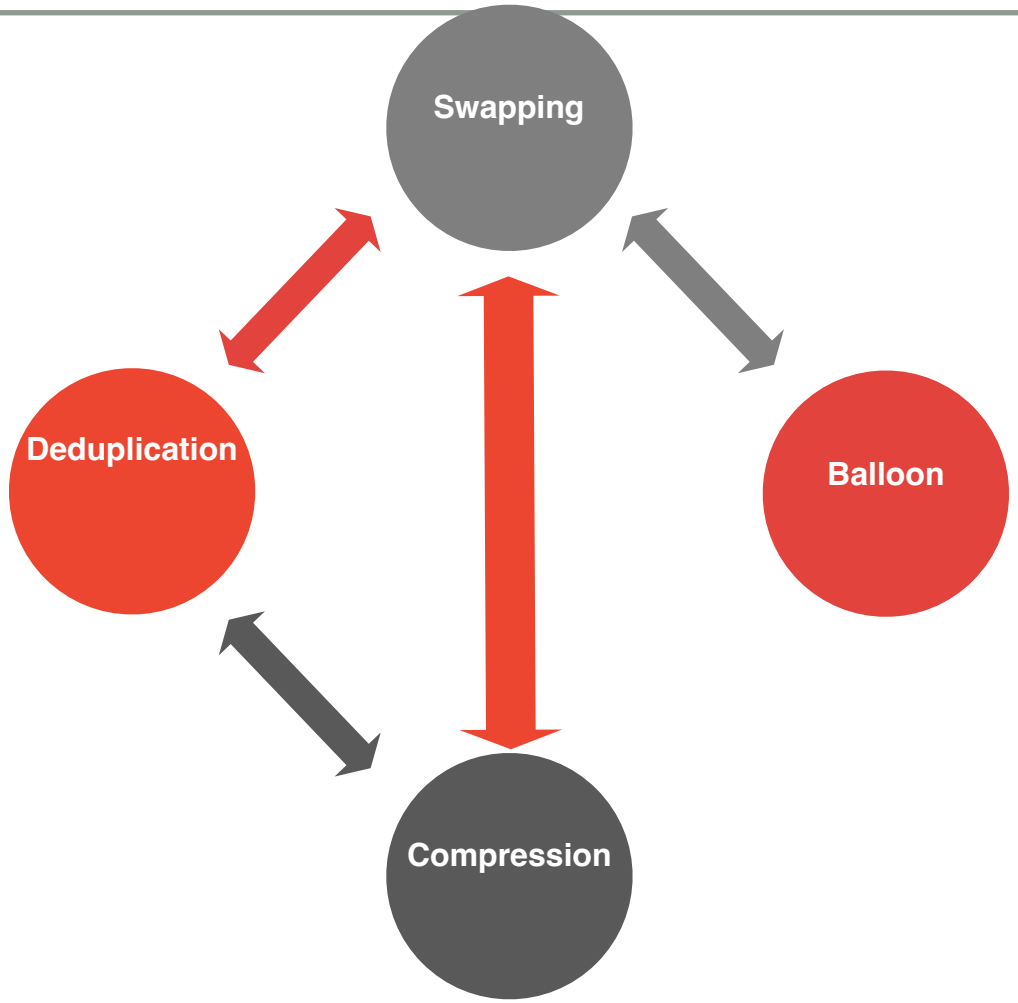
Memory virtualization: the big picture



Balloon + deduplication



Technology interaction



Conclusions

Common resource management task includes quota management, compression, deduplication, replacement + backing store techniques. Combining these solutions is the state of art

Questions?