Automatic ballooning

Luiz Capitulino
lcapitulino@gmail.com
October 22nd, 2013

Agenda

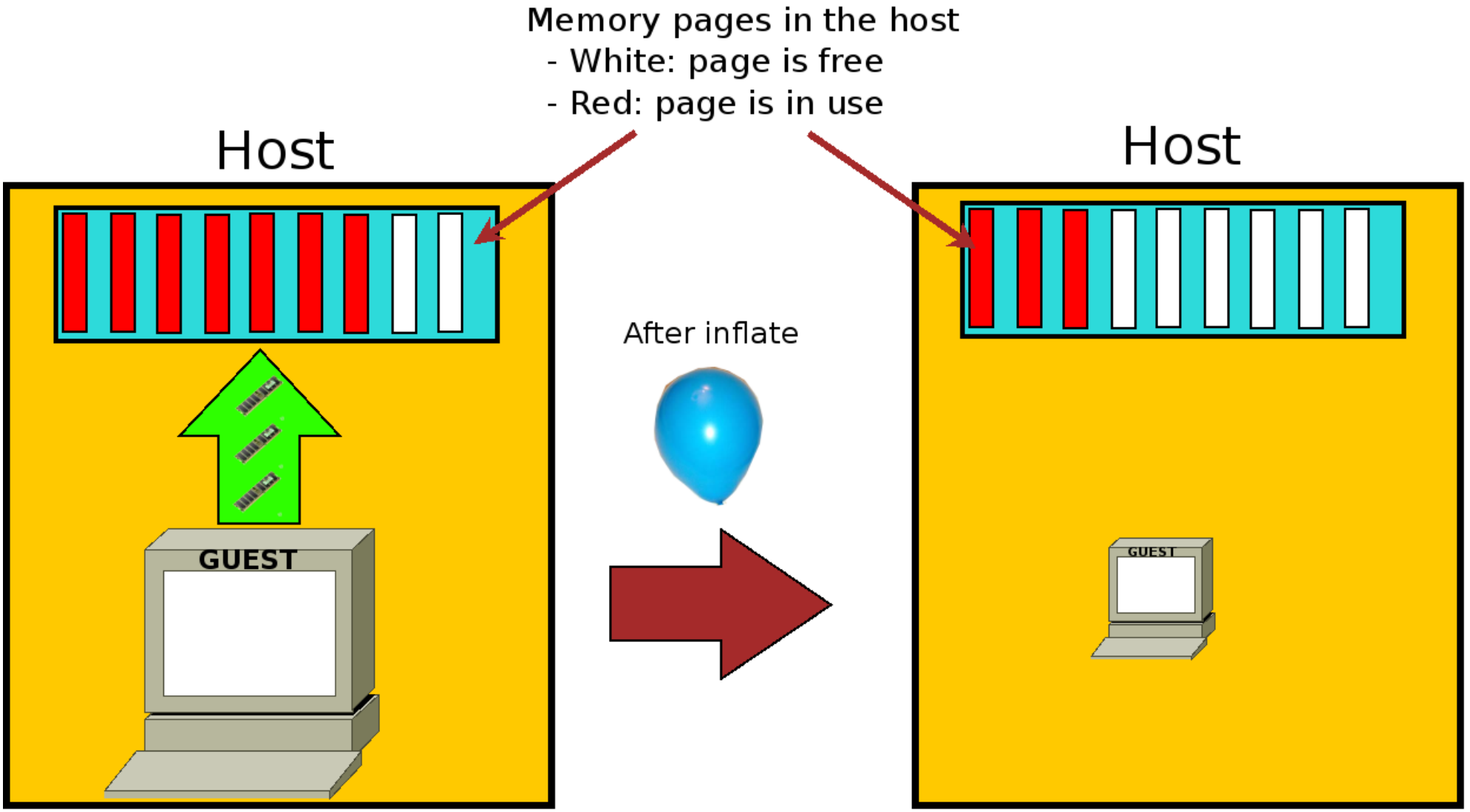
- **The balloon driver**
- **Making it automatic**
- **Testing & some numbers**

The balloon driver

The balloon driver

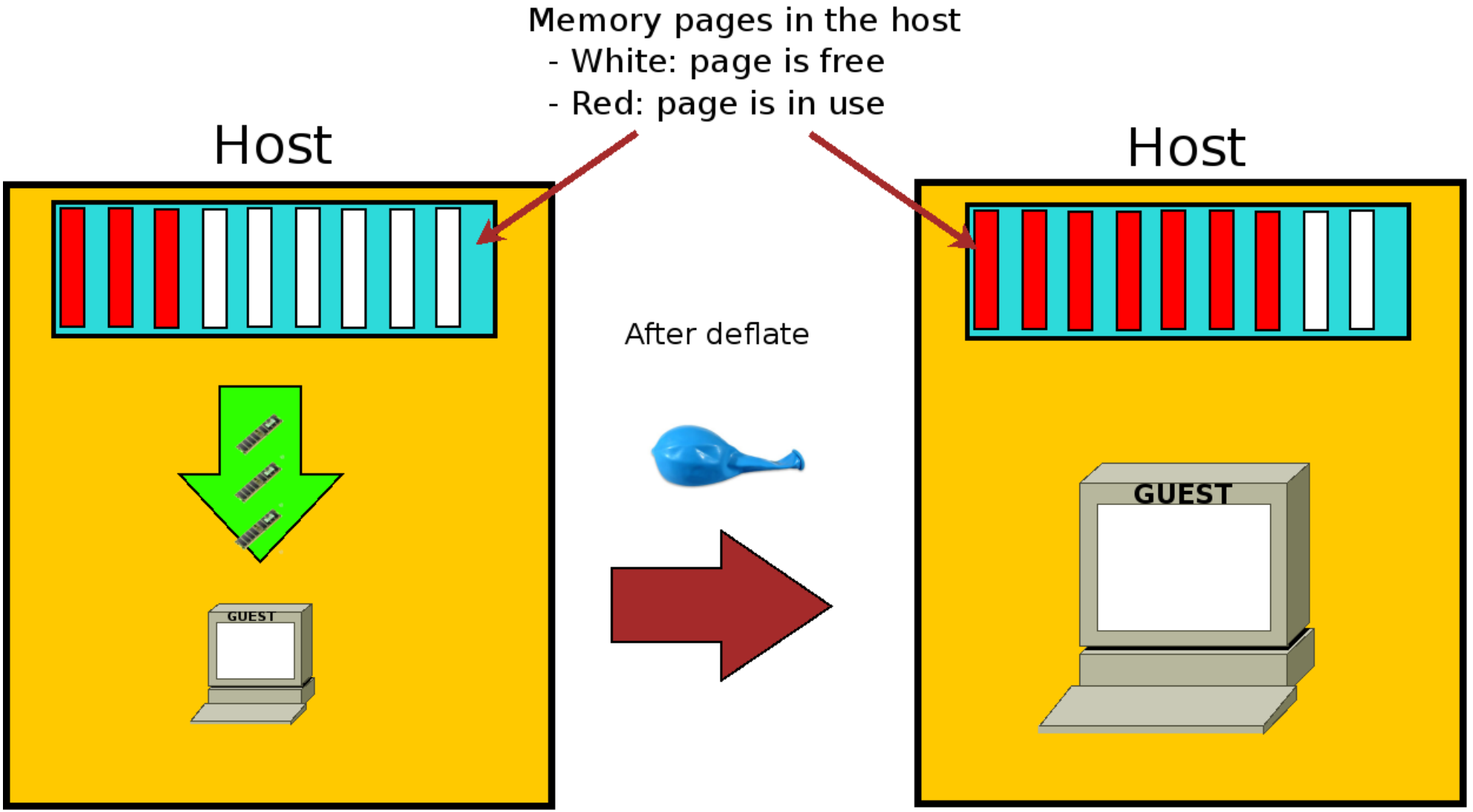
- **Implements two fundamental operations**
  - Inflate: memory is taken from the guest and given to the host
  - Deflate: memory is taken from the host back to the guest
- **Also supports statistics reporting and other features**
- **Available for Linux and Windows guests**

Inflate example



redhat.

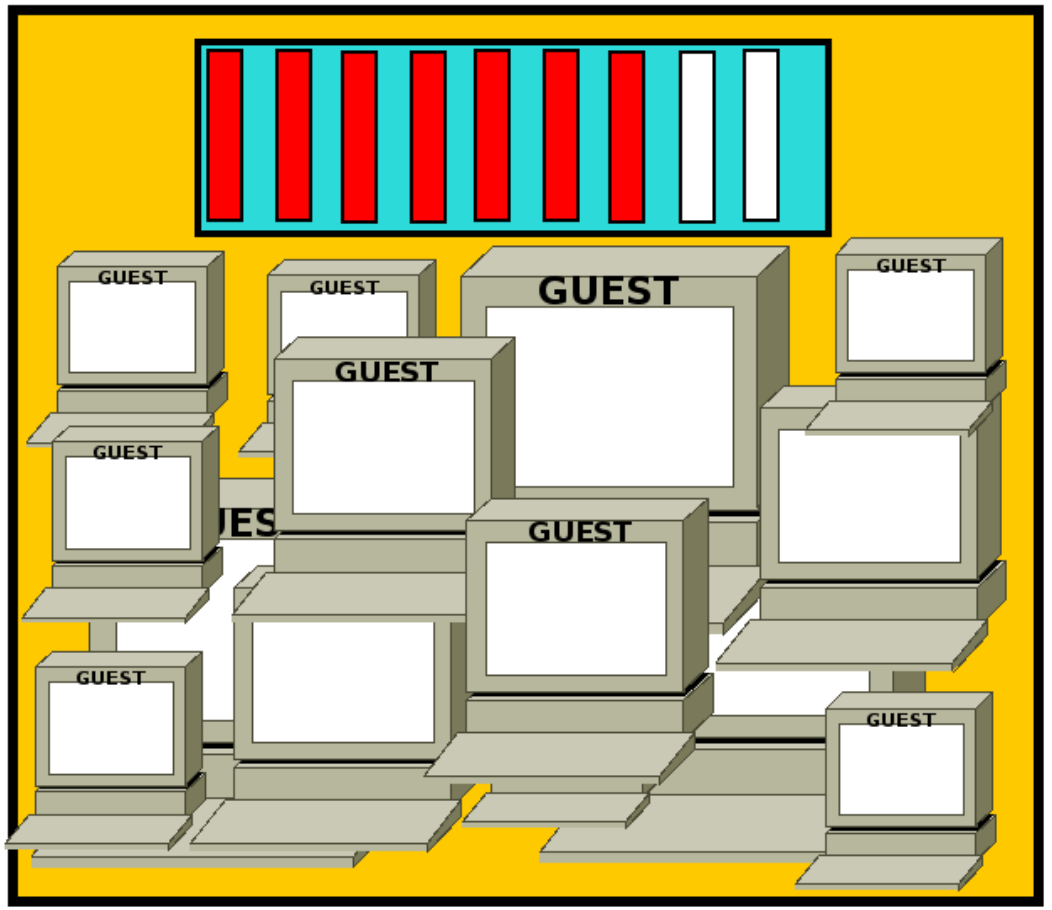
Deflate example



redhat.

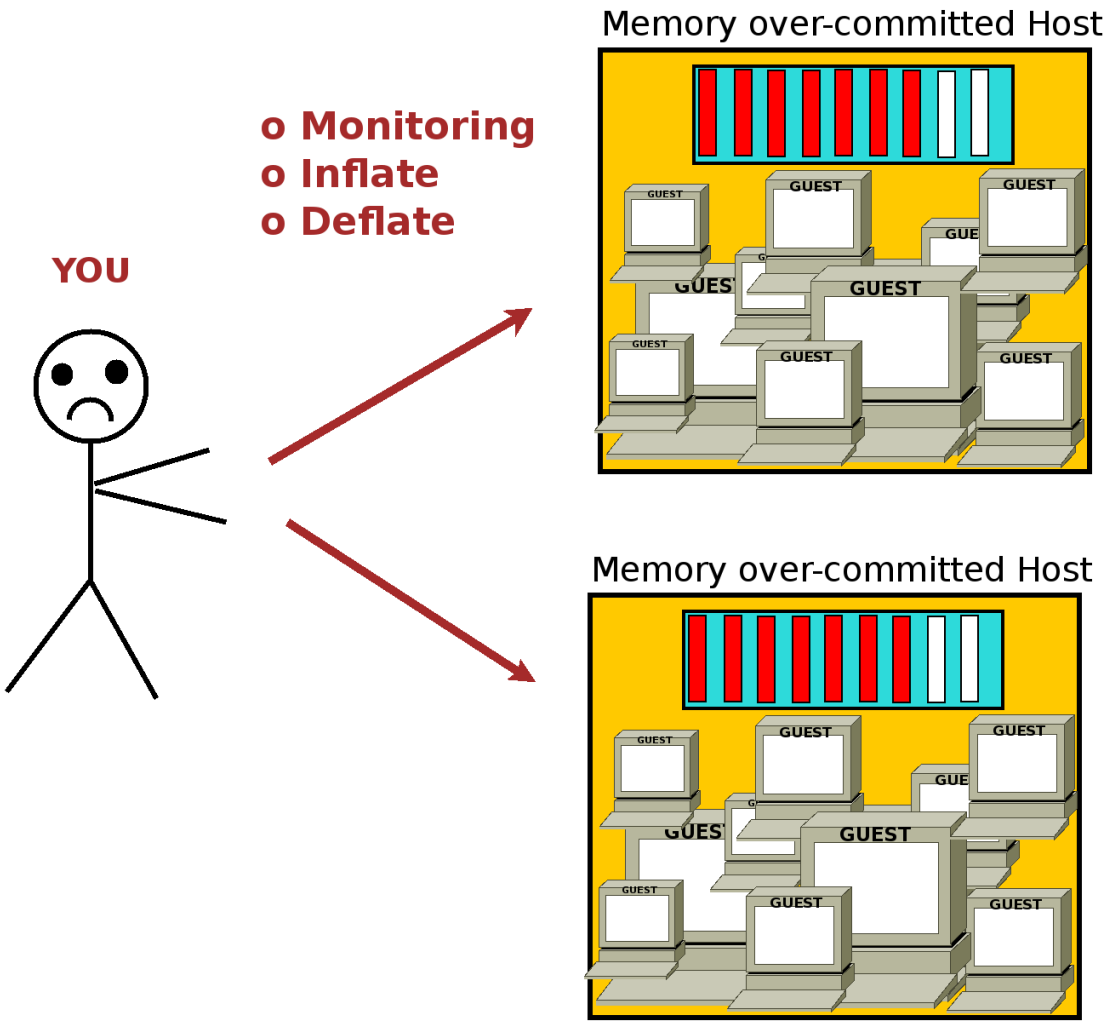
Balloon's primary advantage

Memory over-committed Host



redhat.

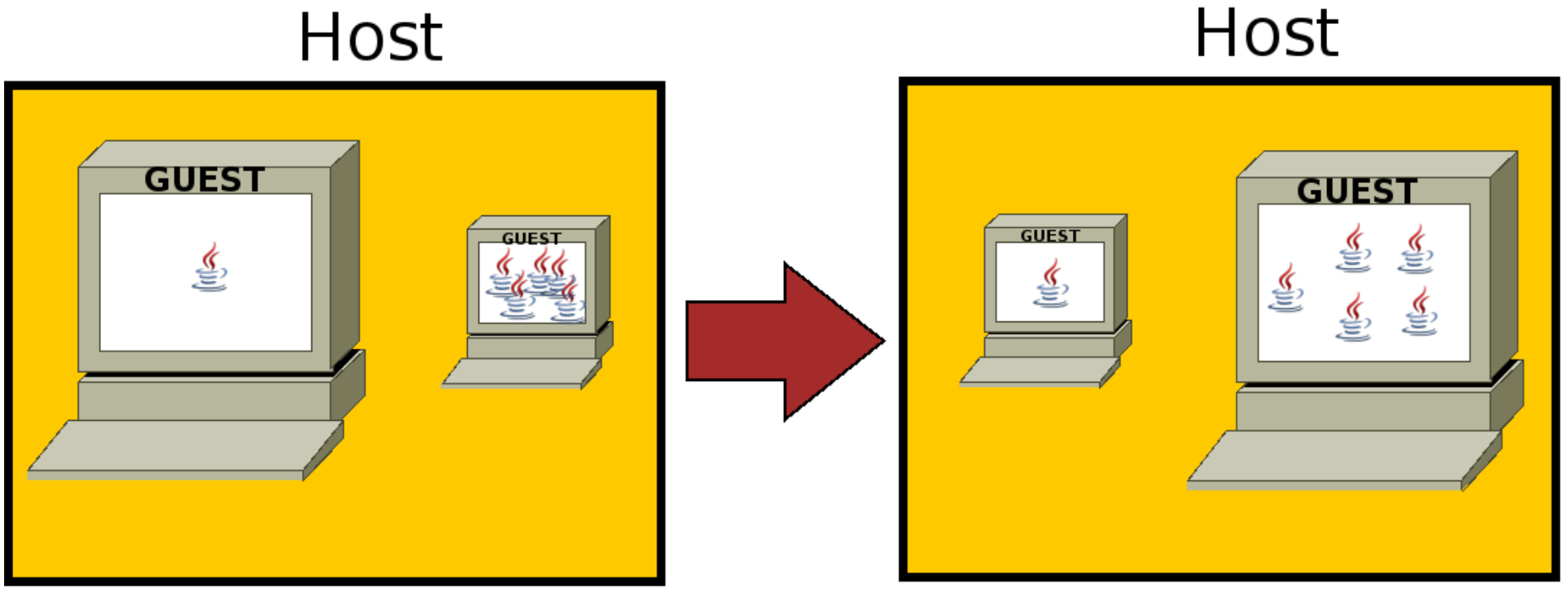
Balloon's primary disadvantage



redhat.

We do have to make it automatic

- Guests automatically shrink on host pressure
- Guests automatically grow when they face pressure themselves
- Guests are automatically managed on memory over-commited Hosts



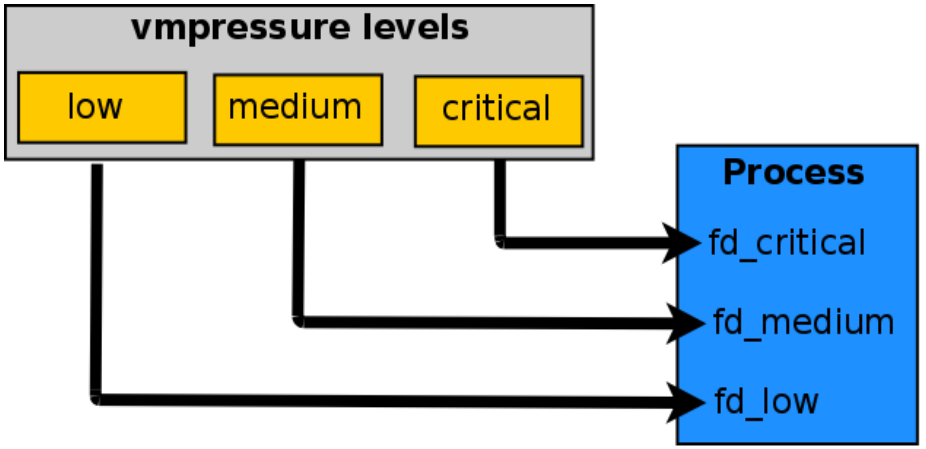
redhat.

Making it automatic

(Based on a design by Rik van Riel, help from Rafael Aquini)

vmpressure events (auto-inflate)

- Added to kernel 3.10 by Anton Vorontsov
- Part of memory controller cgroup
- Defines three pressure levels
  - **LOW:** the kernel is reclaiming memory for new allocations
  - **MEDIUM:** some swapping may be going on
  - **CRITICAL:** the system is thrashing, OOM killer may be on its way to trigger
- User-space is notified via eventfds



vmpressure usage for auto-inflate

- QEMU registers eventfds for low, medium and critical
- QEMU uses pre-defined values to perform auto-inflate
  - Low: 256 (1MB)
  - Medium: 512 (2MB)
  - Critical: 1024 (4MB)
- These values can be run-time tunables

Auto-inflate problems/solutions

- Pre-defined values don't take host nor guest need into consideration
  - **Solution**: the host tells the guest its facing pressure and the guest releases pages accordingly
- QEMU can get as many as 20 events when host is under pressure
  - **Solution**: event throttling in QEMU (1 per sec)
- All event fds are woken up on CRITICAL level
  - **Solution**: demultiplex events in QEMU

shrink callback (auto-deflate)

- Drivers or subsystems can register a function to be called when the kernel is facing memory pressure
- The **guest** virtio-balloon driver implements such a function which performs auto-deflate (ie. memory is reclaimed for the guest)

Auto-deflate problems/solutions

- The shrinker API asks for (only!) 128 pages per call
- Auto-deflate can be delayed due to auto-inflate taking too long

A few words on the current status

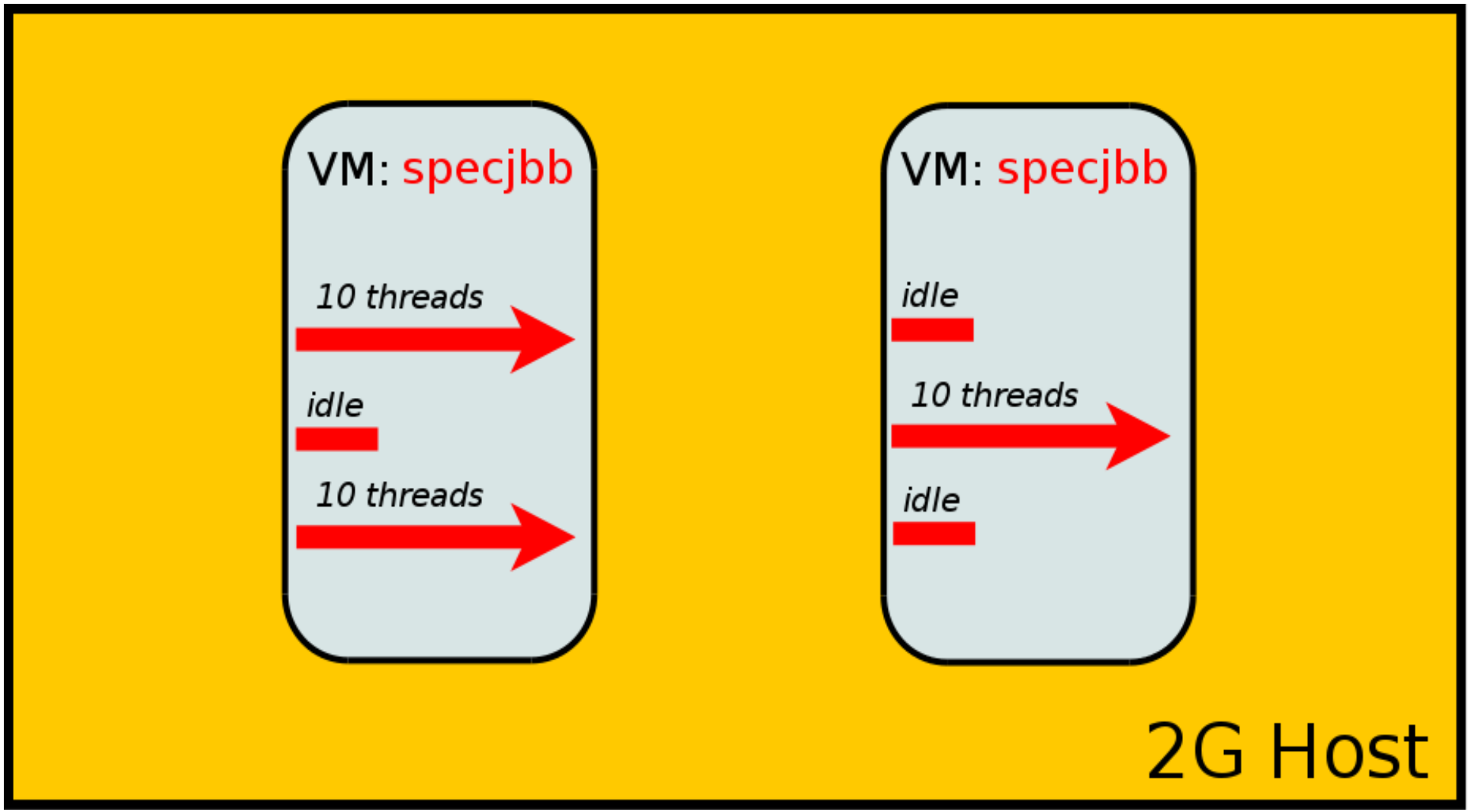
- A prototype exists for almost a year
  - Still pretty experimental
- Two RFC versions posted upstream
  - Need more feedback!

Testing

Take it with a grain of salt

- Very hard to come up with a good test-case
- Smallest change in parameters can change the results
- Several scenarios to be tested

A very simple test-case



Test results: 10 runs average

	Vanilla	auto-ballon	Difference %
Total run time (sec)	441	441	0%
Pages swapped in (host)	46346	41898	-9.60%
Pages swapped out (host)	209710	196080	-6.50%
Specjbb throughput – VMs (bops)	57378.96	58086.61	+1.23%

redhat.


That's all folks!
Luiz Capitulino <lcapitulino@redhat.com>
http://www.linux-kvm.org/Projects/auto-ballooning