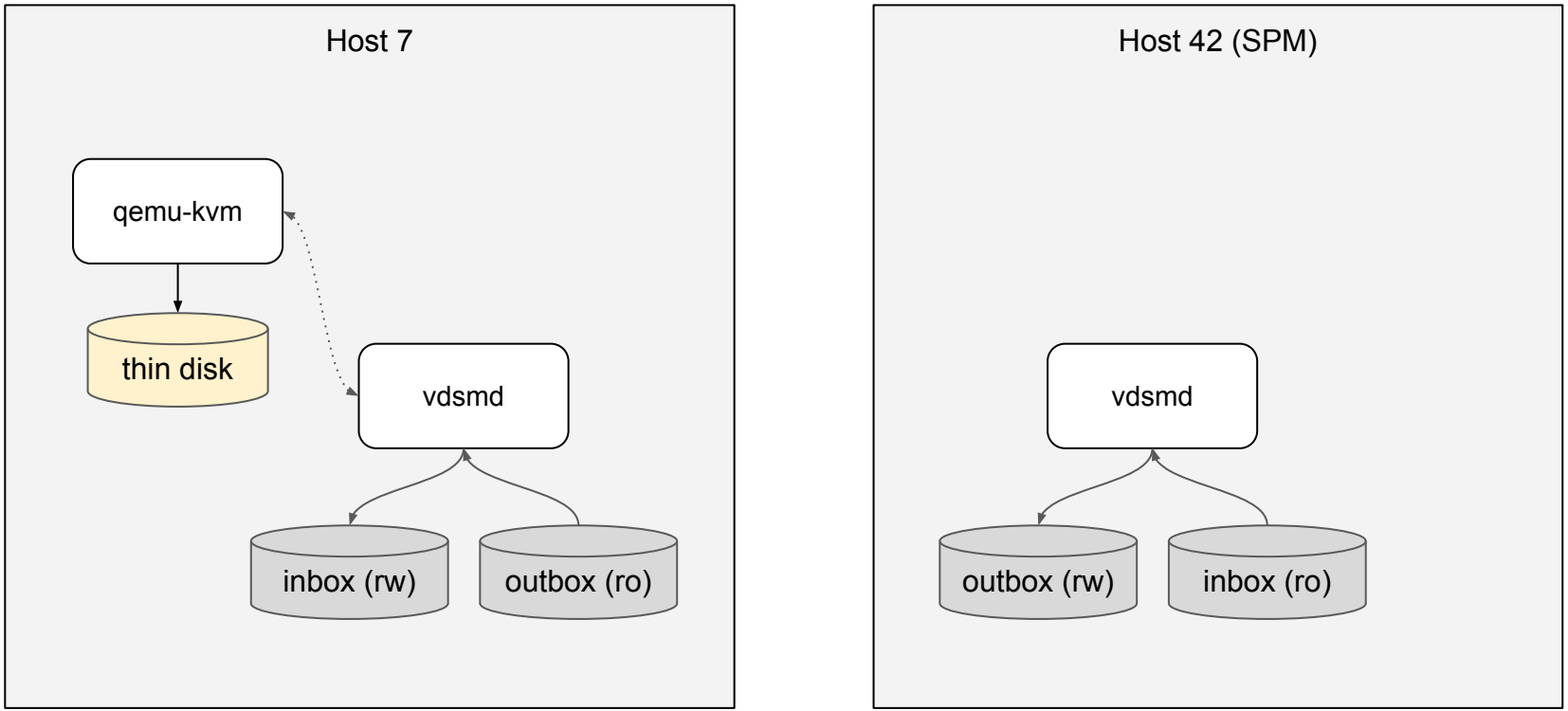
Mailbox Events

Nir Soffer
Principal Software Engineer
nsoffer@redhat.com

RHV Demo April 2022

What is the storage mailbox?


Host 7
qemu-kvm
thin disk
vdsmd
inbox (rw)
outbox (ro)
Host 42 (SPM)
vdsmd
outbox (rw)
inbox (ro)

Extend timeline (average case)

```
00:00.00  Disk becomes too full
00:01.00  Host 7 sends extend mail to the SPM
00:02.00  SPM reads the inbox
00:02.25  SPM extends the disk and sends a reply
00:03.00  Host 7 reads its mailbox
00:03.25  Host 7 refreshes the volume
00:04.00  qemu writes to new area successfully
```

Why mailbox events?

Extend timeline (worst case)

00:00.00 Disk becomes too full

00:02.00 Host 7 sends extend mail to the SPM

00:04.00 SPM reads the inbox

00:04.00 **qemu fails to write with ENOSPC, VM paused**

00:04.25 SPM extends the disk and sends a reply

00:06.00 Host 7 reads its mailbox

00:06.25 Hos 7 refreshes the volume, resume the VM

Storage mailbox is too slow

Why the storage mailbox is slow?

- Every host can write 63 message to its mailbox in the SPM inbox (4 KiB)
- System supports up to 249 hosts
- Reading the SPM inbox reads 1000 KiB from storage
- The inbox can have up to 15,687 messages (typically 0)
- SPM reads the inbox (1000 KiB) every 2 seconds
- Host reads its mailbox (4 KiB) every 2 seconds (when waiting for reply)

How to detect new mail without reading the entire inbox?

You have a new mail!

Inbox events block

There is no host 0 - we use host 0 mailbox for events.

How mailbox events work

- Hosts write an event to the events block when sending mail to SPM
- SPM reads the events block (4 KiB) every 0.5 seconds
- SPM keeps the last event UUID
- If current event is different from last event, SPM reads the inbox (1000 KiB)
- Hosts read their mailbox (4 KiB) every 0.5 seconds to get reply

Extend timeline - with events (average case)

00:00.00 Disk becomes too full

00:01.00 Host 7 sends extend mail to the SPM and writes an event

00:01.25 SPM reads the events block and read the mailbox

00:01.50 SPM extends the disk and sends a reply

00:02.00 Host 7 reads the mailbox

00:02.25 Host 7 refreshes the volume

00:04.00 qemu writes to new area successfully

Extend timeline with events (worst case)

```
00:00.00  Disk becomes too full
00:02.00  Host 7 sends extend mail to SPM and writes an event
00:02.50  SPM reads the events block and read the mailbox
00:02.75  SPM extends the disk and sends a reply
00:03.00  Host 7 reads the mailbox
00:03.25  Host 7 refreshes the volume
00:04.00  qemu writes to new area successfully
```

Handling concurrent events

00:00.00 SPM last event: 1f32b266-42e1-413b-9e7e-6f66a34873cd

00:00.23 Host 7 writes event: 8086a64c-d020-49e7-8da1-0d7e032513a6

Host 42 writes event: 4e5b3a2d-4303-4ff6-9461-be25025bf47e

00:00.34 Host 3 writes event: 0b3ca27d-a92c-4ce0-80d1-857caf719cdc

00:00.50 SPM reads event: 0b3ca27d-a92c-4ce0-80d1-857caf719cdc

reads entire inbox, process all mail

Environment with old and new hosts

- Old hosts do not write mailbox events, and read mailbox every 2 seconds
- New hosts write mailbox events and read mailbox every 0.5 seconds
- Old SPM does not monitor mailbox events, serving new hosts slower (like old hosts)
- New SPM monitors mailbox events, serving new hosts faster, and old hosts slower
- Works same or better than an old environment

How fast can we write without pausing?

version	chunk size	utilization	event interval	write rate
oVirt 4.4	1 GiB	50%	-	**75 MiB/s**
oVirt 4.5.0 alpha	2.5 GiB	20%	-	**350 MiB/s**
oVirt 4.5.0 beta	2.5 GiB	20%	0.50	**650 MiB/s[1]**
oVirt 4.5.1	2.5 GiB	20%	0.50	**1300 MiB/s[2]**
oVirt 4.5.1	2.5 GiB	20%	0.25	**1400 MiB/s[2]**

[1] At 700 MiB/s VM paused once when extending disk 20 times.
[2] Using https://github.com/oVirt/vdsm/pull/124.

Disabling mailbox events

If mailbox events cause trouble, they can be disabled:

```
$ cat /etc/vdsm/vdsm.conf.d/99-local.conf
[mailbox]
events_enable = false
```

The drop in file should be installed on all hosts.

When events are disabled, hosts do not write events or monitor their mailbox using the event interval, and the SPM does not read events.

Configurable event interval

The event interval can be modified:

```
$ cat /etc/vdsm/vdsm.conf.d/99-local.conf
[mailbox]
events_interval = 0.25
```

Testing shows extend time of 0.6 ± 0.3 seconds with this configuration.

Upstream, will be available in next oVirt 4.5.0 beta build.

Can be the default in oVit 4.5.1.

Related changes

Remove the 0-2 seconds delay before sending the extend mail

```
00:00.00  Disk becomes too full
00:02.00  Host 7 sends extend mail to the SPM and writes an event
```

Tracked in https://github.com/oVirt/vdsm/issues/85

PR: https://github.com/oVirt/vdsm/pull/124

Will be available in oVirt 4.5.1.

Future work

If all hosts support mailbox events, the SPM does not need to read the mailbox every 2 seconds.

Hosts report new capability:

```
"mailbox_events": true
```

If all hosts report this capability, engine can configure the SPM to use only mailbox events.

Minimize I/O and CPU usage on SPM host.

Tracked in https://github.com/oVirt/vdsm/issues/175

More info

- Minimize storage mailbox latency
  https://github.com/oVirt/vdsm/issues/102
- mailbox: Minimize messages latency
  https://github.com/oVirt/vdsm/pull/103
- mailbox: Configurable event interval
  https://github.com/oVirt/vdsm/pull/110
- [RFE] Default thin provisioning extension thresholds should match modern hardware
  https://bugzilla.redhat.com/2051997

Questions?