Failure handling for SvMotion

- by Anupama Chandwani

Declaring failure

- VMX detects a failure: svmFailure = TRUE
- Pre mirror installation:

  SVMotionMirrorModeThread(): set svmFailure and exit

  SVMotionMirrorModeThreadDone()

      SVMotionCleanupCB()

          SVMotionSetFailure() or SVMotionSetSuccess()
- Post mirror installation:

  SVMotionMirrorModeThread(): set svmFailure and exit

      Checkpoint_Stun(): callback SVMotionStunForCleanupCB()

  SVMotionMirrorModeThreadDone()

      SVMotionCleanupCB()

          SVMotionSetFailure() or SVMotionSetSuccess()

Failure code flow: VMX

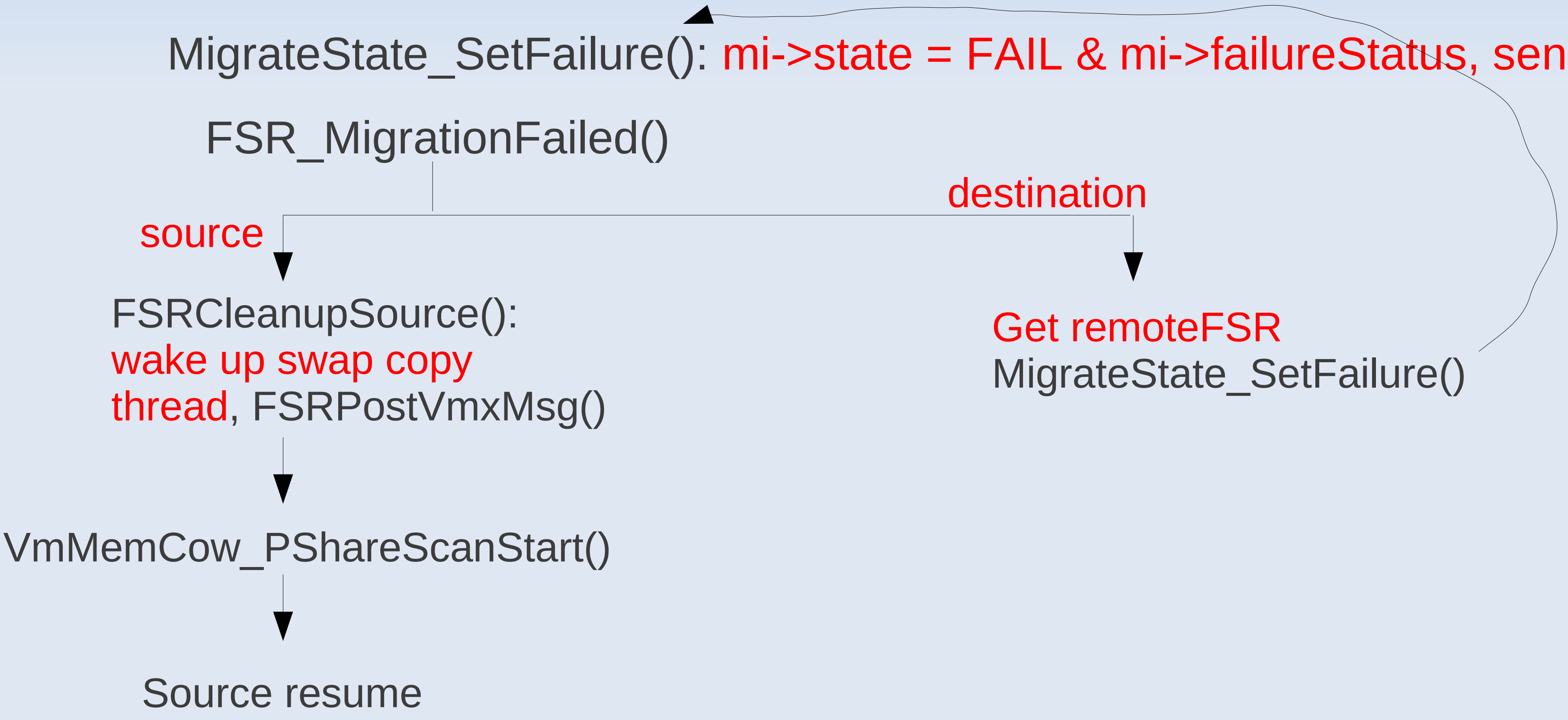
```
SVMotionCleanupCB()
  SVMotionSetFailure()
    Migrate_SetFailureMsgList()
      - migrationState.failureCode = ERROR
      - MigratePlatformSetFailure() - conti.. nxt slide
      - Fire MIGRATE_EVENT_SET_FAILURE vFC & FT registered callbacks
      - SVMotion_Cleanup(): BULL should be held
        - if XvMotion : flush all IOs, Disk_CloseAll() and free files & disks linked lists.
        - Signal all semaphores
        - schedule CleanupGroup in workerQueue – conti.. two slides later
        - set phase ”SVMPhase_Cleanup”
```

Failure code flow: VMKernel

MigratePlatformSetFailure() : Accept VOB

VMKernel_MigrationFailure()

Migrate_VMXMigrationFailure()



MigrateState_SetFailure(): mi->state = FAIL & mi->failureStatus, send VOBs

FSR_MigrationFailed()

source

FSRCleanupSource():
wake up swap copy
thread, FSRPostVmxMsg()

VmMemCow_PShareScanStart()

Source resume

destination

Get remoteFSR
MigrateState_SetFailure()

Failure code flow: VMKernel

FSR_CleanupMigration:

Set ”resumeState” for source to be able to resume.

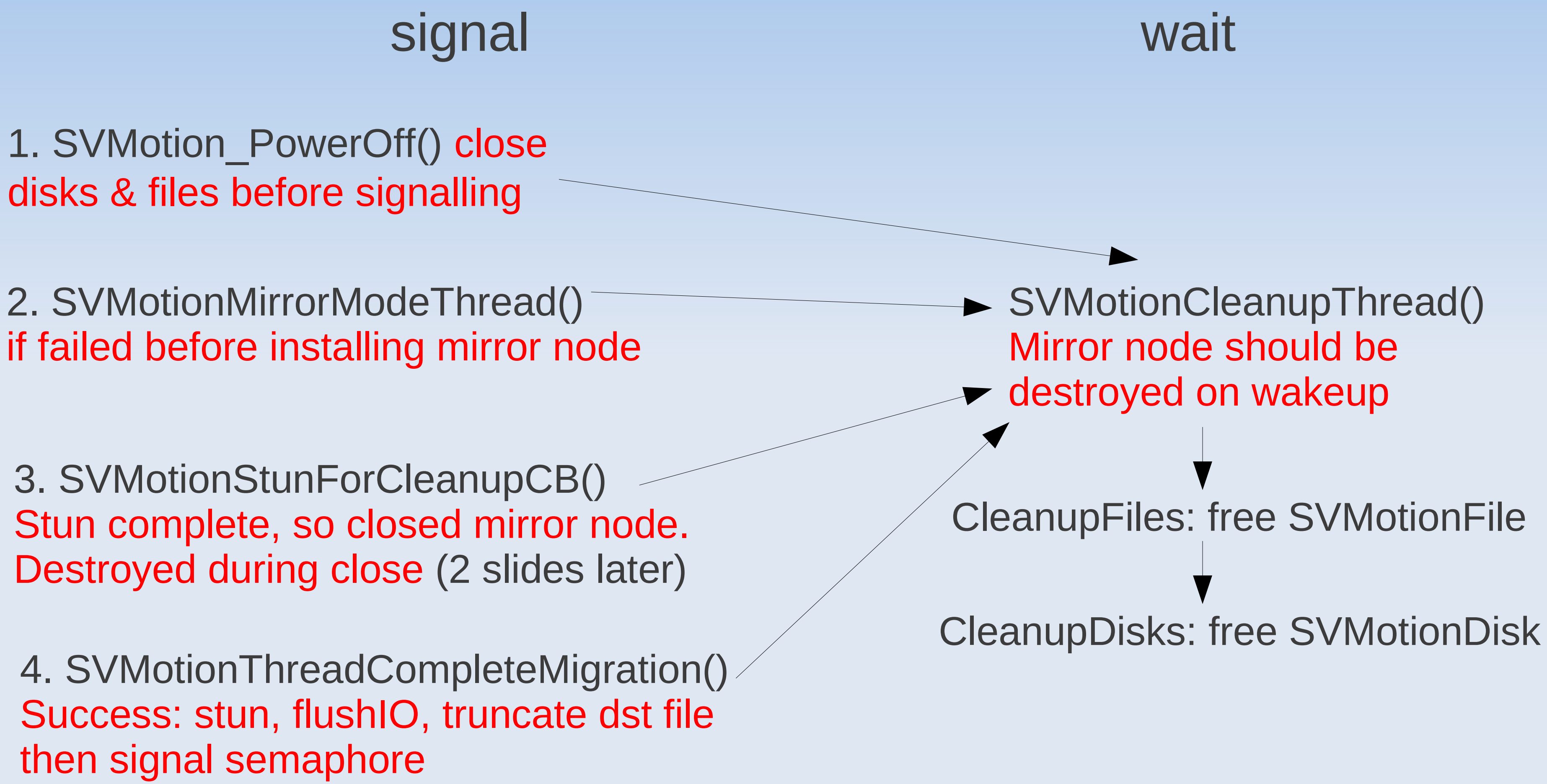
- FSRResumeState: Automic variable to select victor for resuming
- FSR_RESUME_NONE: Anyone can resume
- FSR_RESUME_SOURCE: Source is in the process of resuming.
- FSR_RESUME_DEST: Destination is in the process of resuming

Synchronization

- Following functions execute in parallel
  - SVMotionCleanupThread() / SVMotion_Cleanup()
  - Checkpoint_Stun()

    SVMotionStunForCleanupCB() - Close all disks and files, unstun.
- Sync done by cleanupSemaphore.

cleanupSemaphore



signal

wait

1. SVMotion_PowerOff() close disks & files before signalling

2. SVMotionMirrorModeThread()
if failed before installing mirror node

3. SVMotionStunForCleanupCB()
Stun complete, so closed mirror node.
Destroyed during close (2 slides later)

4. SVMotionThreadCompleteMigration()
Success: stun, flushIO, truncate dst file
then signal semaphore

SVMotionCleanupThread()
Mirror node should be
destroyed on wakeup

CleanupFiles: free SVMotionFile

CleanupDisks: free SVMotionDisk

svMotionCleanupGroup

SVMotionCleanupThread()

- Wait for copy bitmap group to complete

- If copy thread scheduled, wait for svmThreadDone: copy thread to complete

- Wait for final stun/unstun (remove mirror node): cleanupSemaphore

- CleanupFiles, CleanupDisks: free files/disks linkedlists, close dest file/disk

- Destroy all semaphores

- Set phase as ”SVMPhase_NULL”. Important for SVMotion_PowerOff() (2 slides later)

- destroy cleanupSemaphore

Destroy mirror node

Places where source and destination fds are closed:

- SVMotion_DiskCloseCB: for disks & digest disks
- SVMotion_PowerOff(), MigrateStunCallback(), SVMotionStunForCleanupCB() for files

  SVMotion_CloseSourceFiles()

  - filecopyOpsTable close source file callback
  - destroy mirror node
- Destination disk/file is closed in SVMotionCleanupThread() while freeing SVMotionFile and SVMotionDisk linkedlists.

VM PowerOff

SVMotion_PowerOff(): need to destroy mirror nodes

- if phase == SVMPhase_NULL, means svMotionCleanupGroup already executed.

free svmotionGroups and exit. Wait for cleanupGroup to complete.

- else: wait for bitmap and disk copy groups to complete

- Disk_CloseAll() on source disks: mirror node still installed, so this closes mirror node. Mirror nodes destroyed in SVMotion_DiskCloseCB.

- SVMotion_CloseSourceFiles(): Close mirror node on source

- SVMotionCleanupThread waits for final stun & unstun on source to resume. But in powerOff case, source does not resume, so signal cleanupSemaphore

- Waiting for cleanupSemaphore in SVMotionCleanupThread before CleanupDisks to make sure mirror node is closed here.

TODO: SVMotion cleanup

- To cleanup the svmotion cleanup code, we need a state machine with associated callbacks.
- The following functions will each advance the state machine based on it's previous state and call callbacks associated to the state they transition the state machine to.

  SVMotionCleanupThread(), SVMotion_Cleanup(), SVMotion_PowerOff(), Checkpoint_Stun(), SVMotionStunForCleanupCB()
- For eg:
  - destroyMirrorNode is a state followed by closeSourceDiskFile.
  - Stun and Unstun should also be states and code should not wait on semaphores to wakeup from a stun/unstun completion callback.