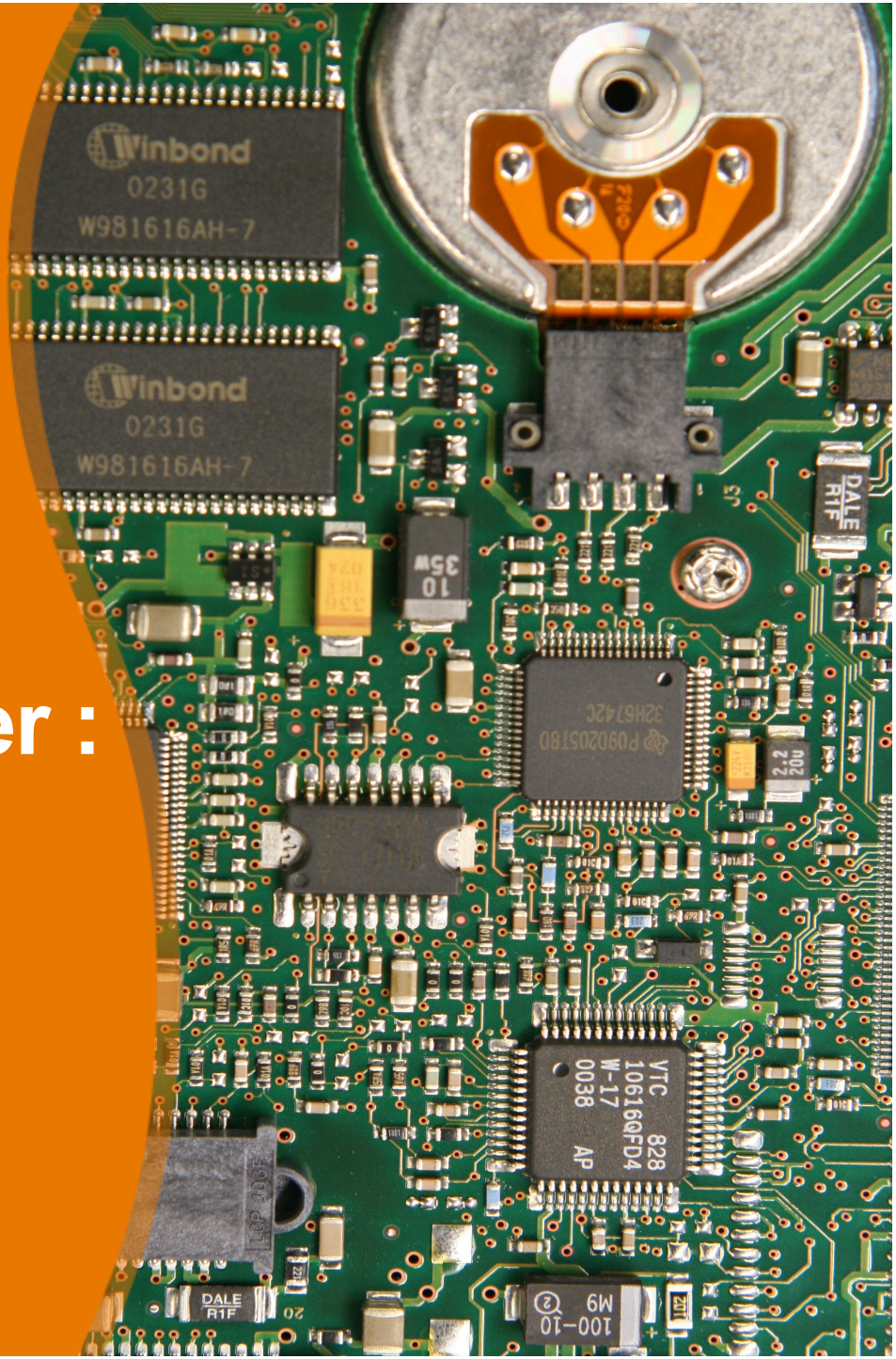
Solaris Volume Manager : Disk Error Handling



Disk Error Handling

- Different forms of disk error
  - > Media errors
  - > Selection errors
- How SVM reacts
  - > Stripes, Concats & Soft partitions
  - > Mirrors
  - > RAID-5

Media Errors

- Happen only for a read / write command
- Can be from :
  - > Failed command on the transport path
  - > Lost command on the device
  - > Physical media defect
- Retries carried out by sd driver

`

Selection Errors

- Happen for any SCSI command
- Can be from :
  - > Power failure to the device
  - > HBA failure
  - > Physical transport defect
- Retries carried out by sd driver

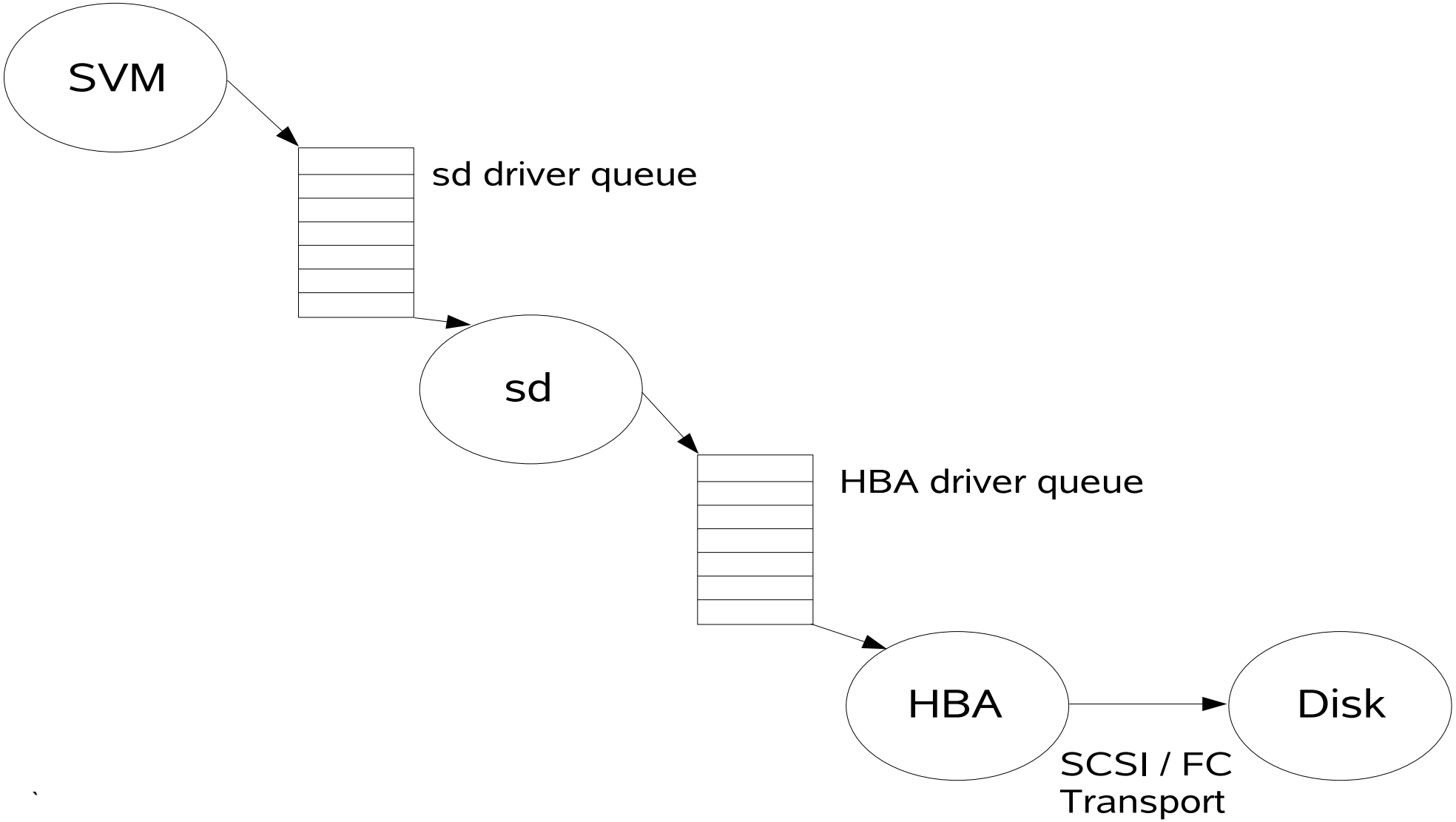
Driver Retries

- Pre-configured number of retries
- Specific timeouts between retries
- Tunable in both sd and HBA drivers

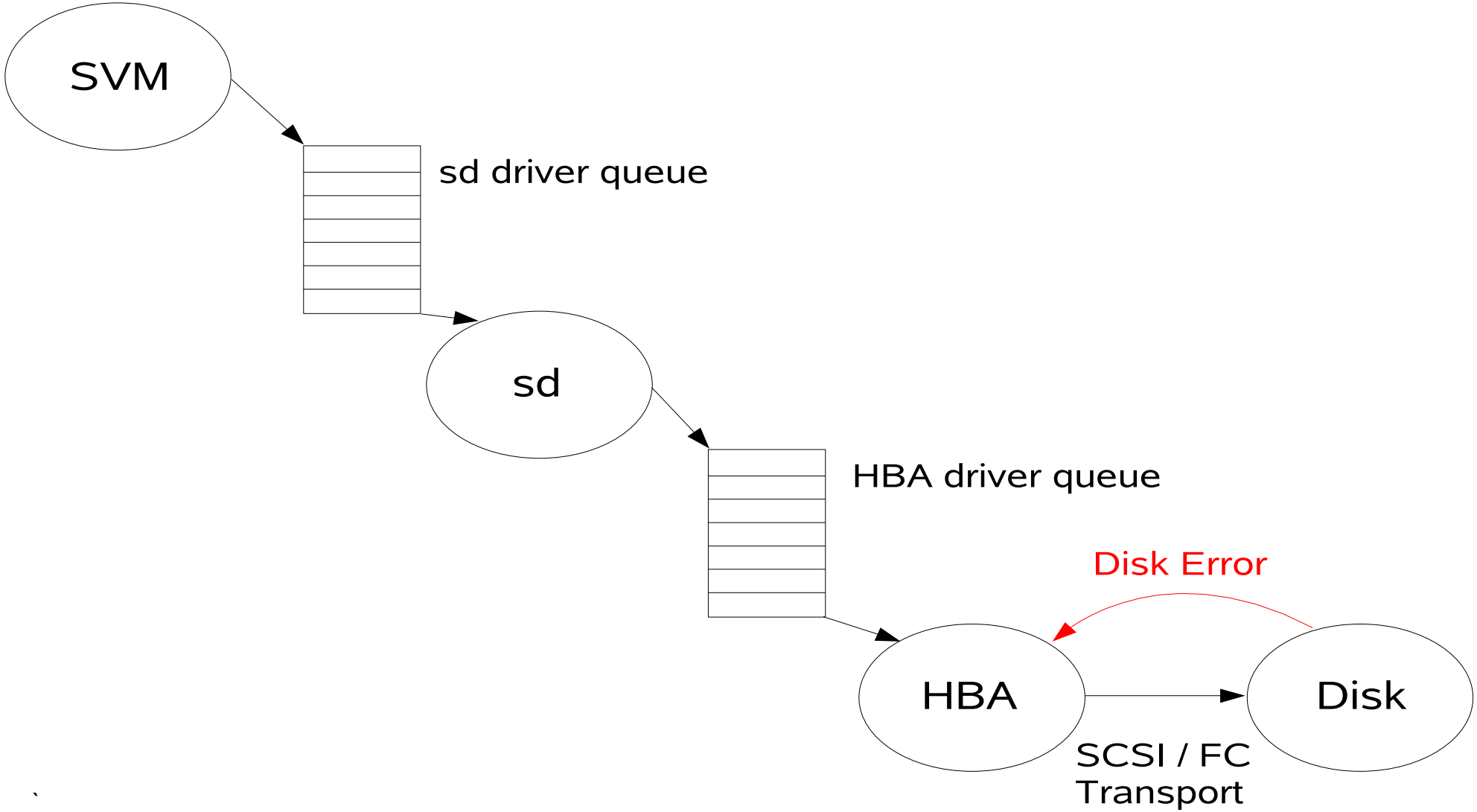
```
# cat /kernel/drv/qus.conf
scsi-selection-timeout=4;
scsi_reset_delay=500;

# tail /etc/system
set sd:sd_retry_count=0x2
set sd:sd_io_time=0x10
set sd:sd_error_level=0x0
```

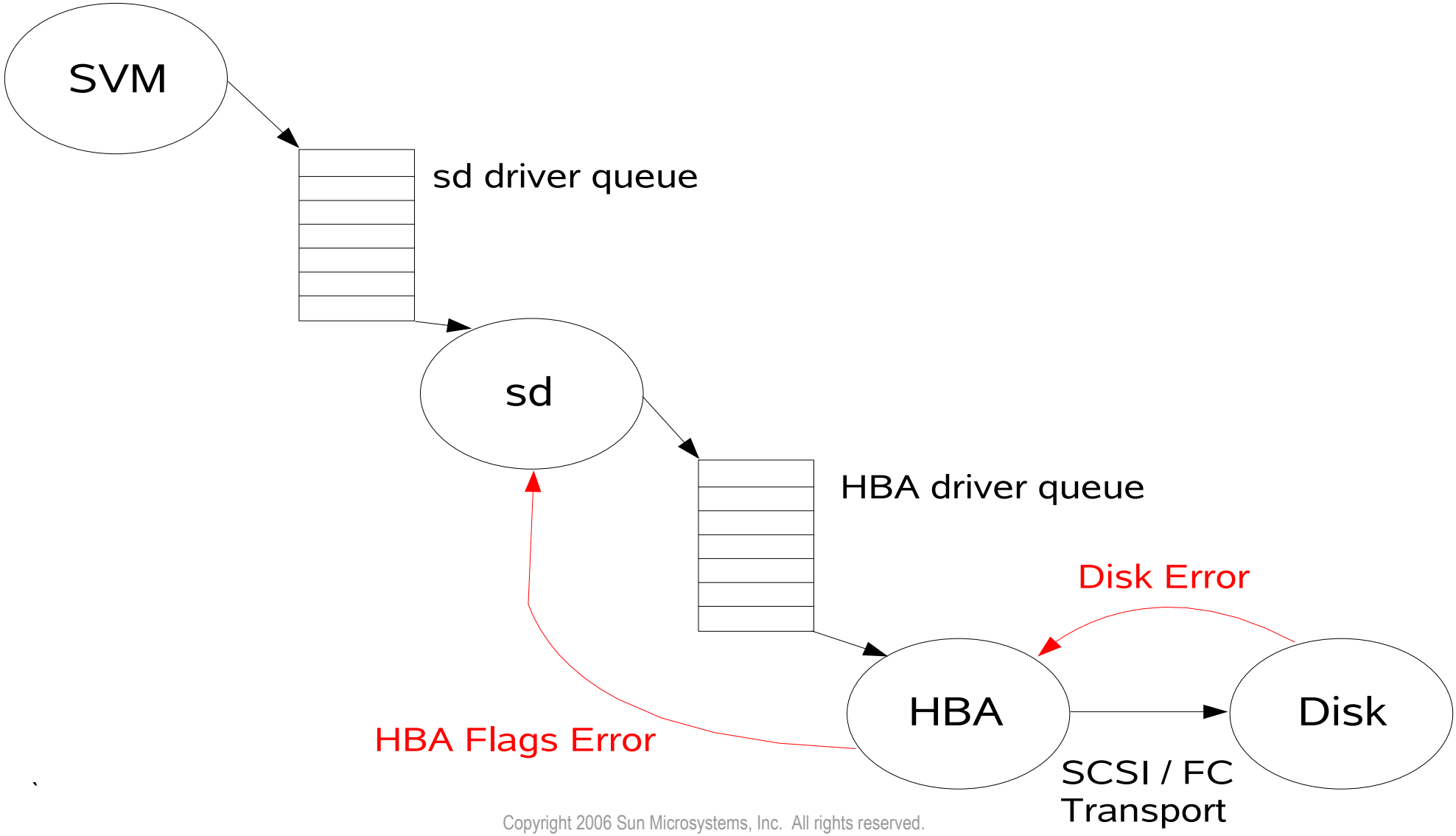
Driver Retries – I/O Issued



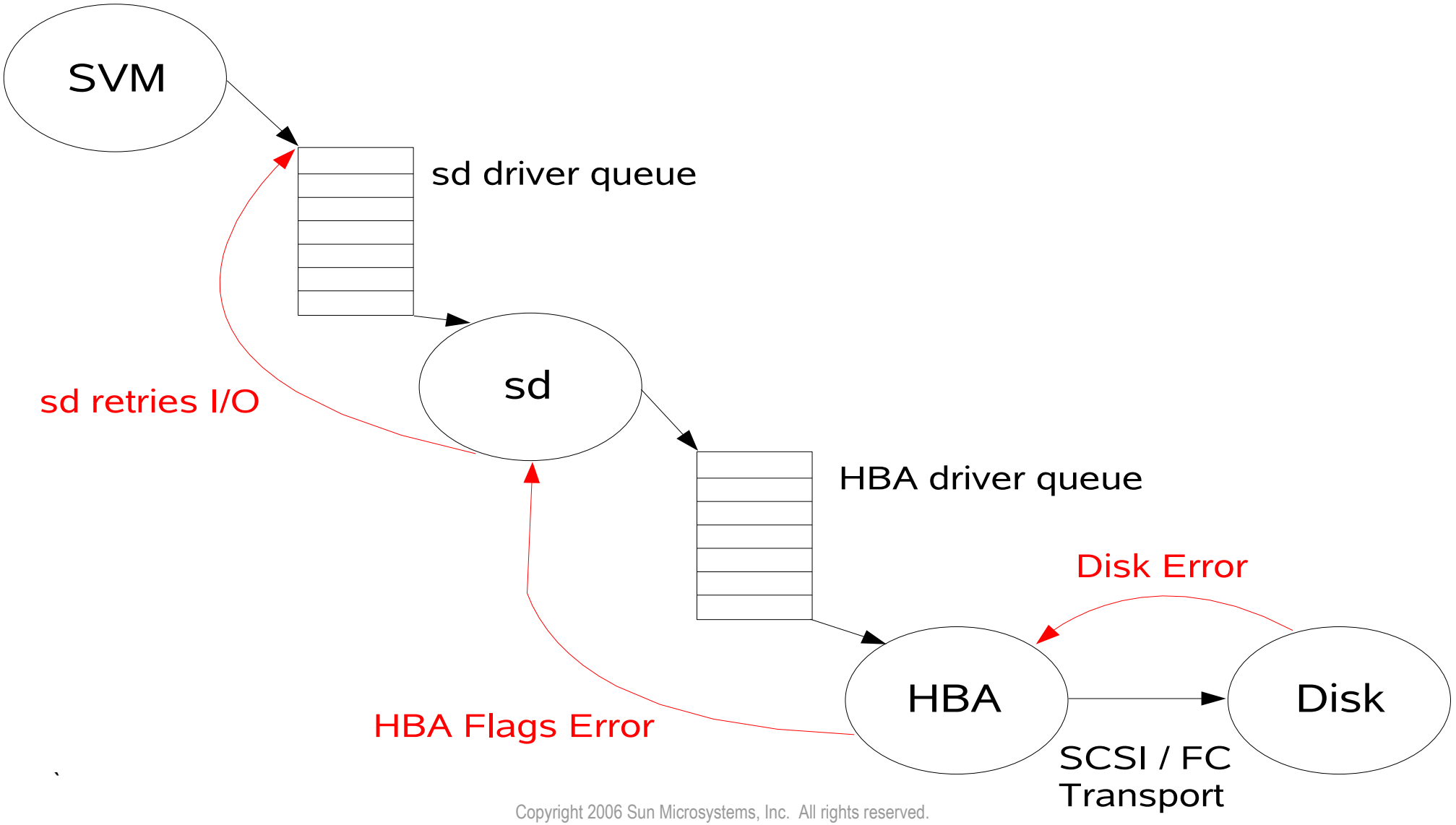
Driver Retries – I/O Fails



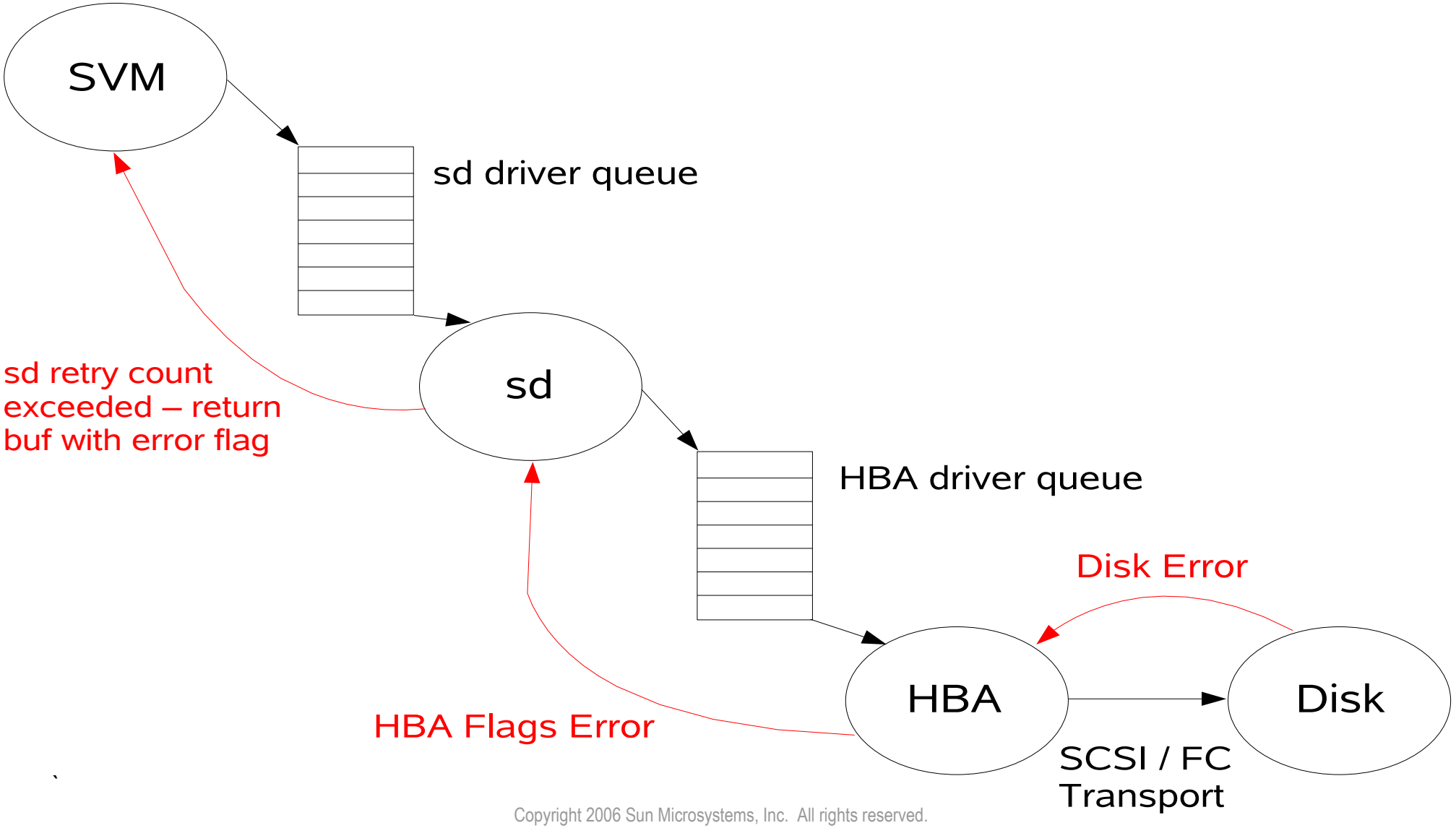
Driver Retries – HBA -> sd drivers



Driver Retries – I/O Re-Queued



Driver Retries – Retries Failed



Driver Retries

- Can be very slow
  - > Each retry for a selection timeout can be 60 seconds
  - > Each retry goes to be back of the queue
  - > Other I/O's ahead also each take 60 seconds to fail
- No failure back to SVM until retries are exhausted
- Tuning very worthwhile
  - > Need to understand the config concerned

`

SVM Error Handling

- Only kicks in when a buf returns with an error
- Always results in SVM read / write error on a device
  - md_stripe: WARNING: md: d20: write error on /dev/dsk/c1t1d0s3
- Action taken depends on metadevice type

Stripe / Concat / SP Errors

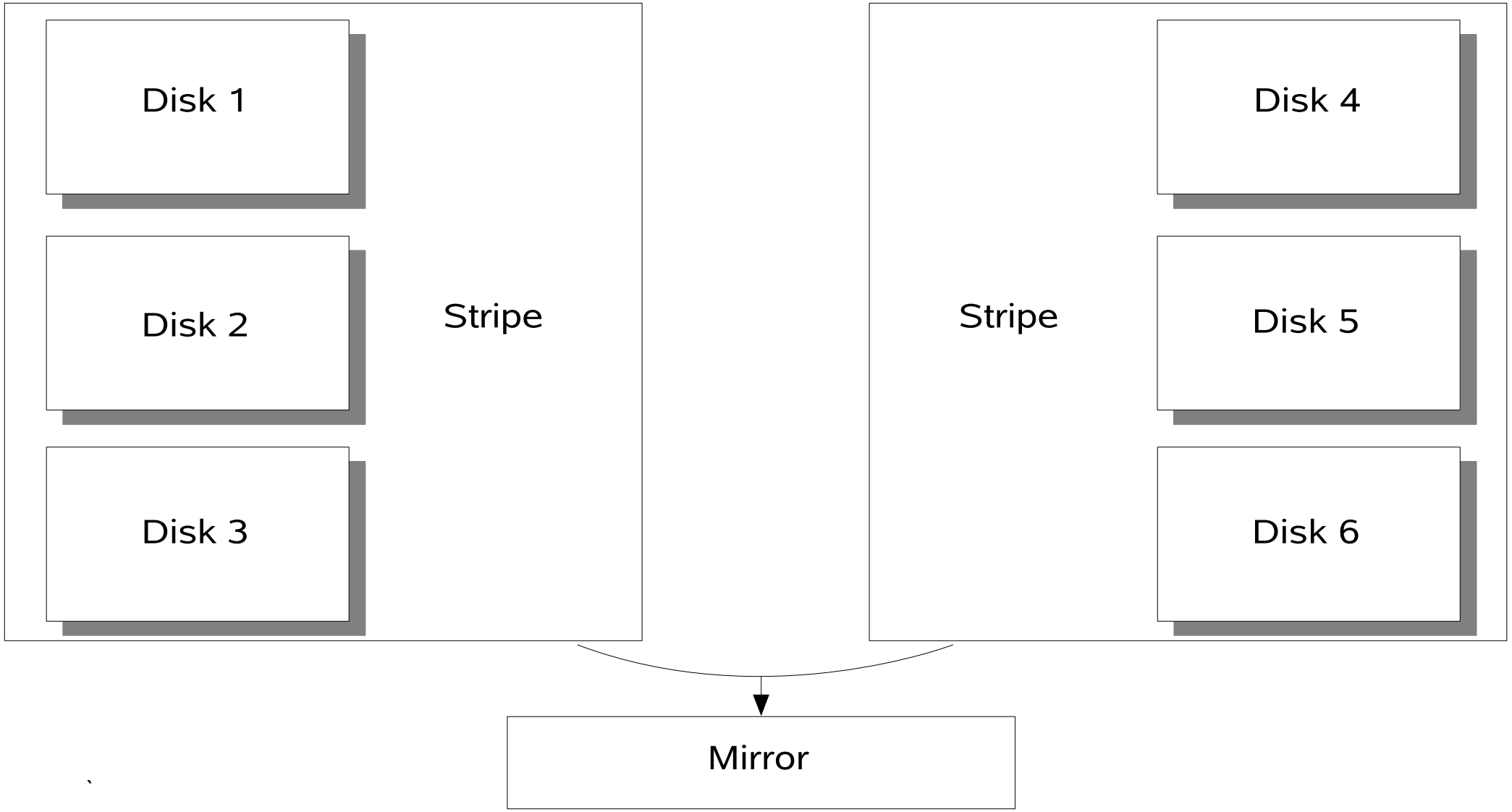
- No redundancy – no action can be taken
- Failures treated as unrecoverable
- Buf error passed on to the calling layer
- Any fault in these devices loses all data on the device

Mirror Errors

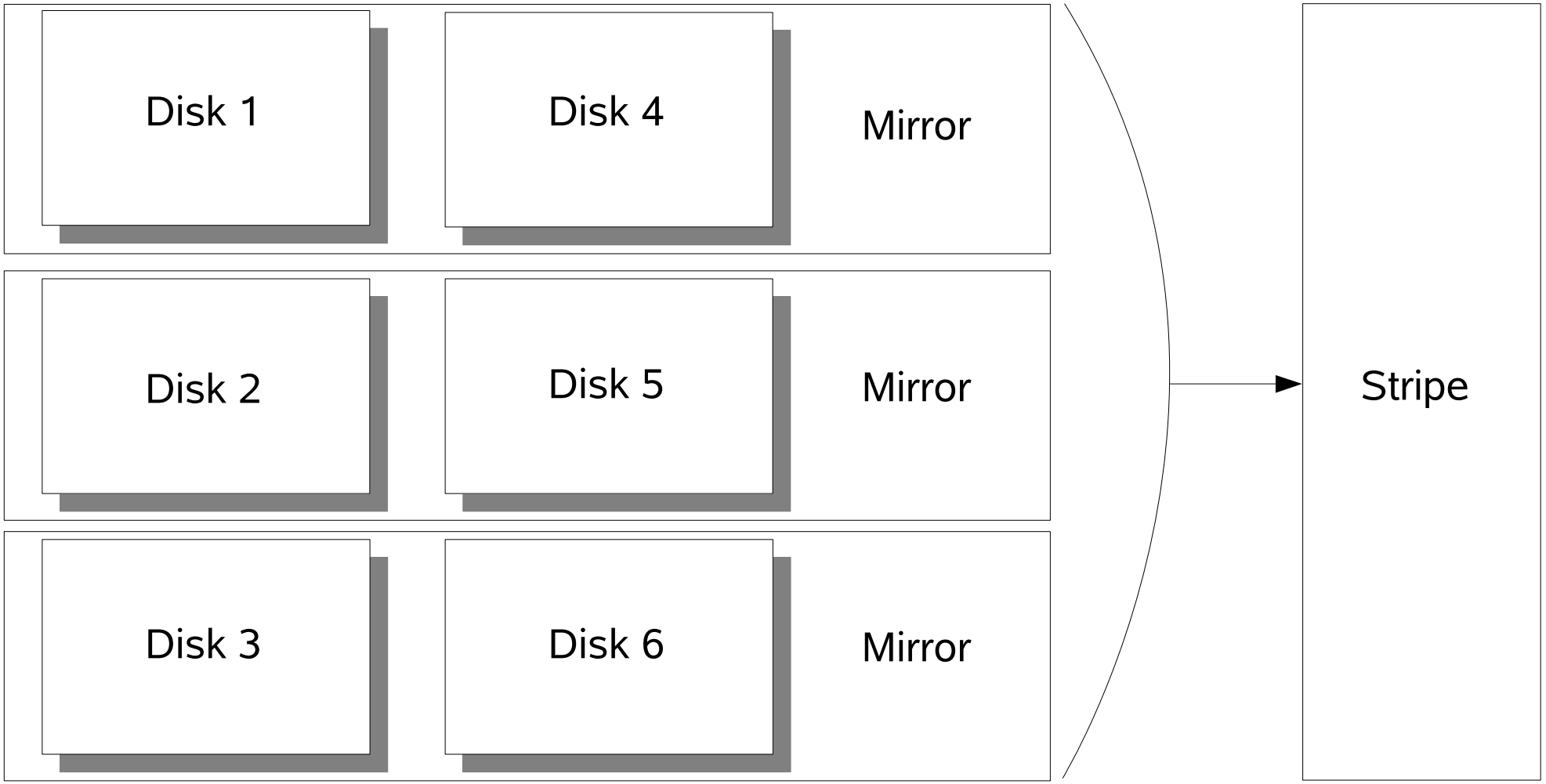
- Error on a sub-mirror
- Data preserved
- May survive multiple-disk failures
  - > Raid 0+1 – loses whole sub-mirror on single failure
  - > Raid 1+0 – only loses failed disk
  - > SVM always uses Raid 1+0 where possible
- State model used to determine action

`

RAID 0 + 1



RAID 1 + 0

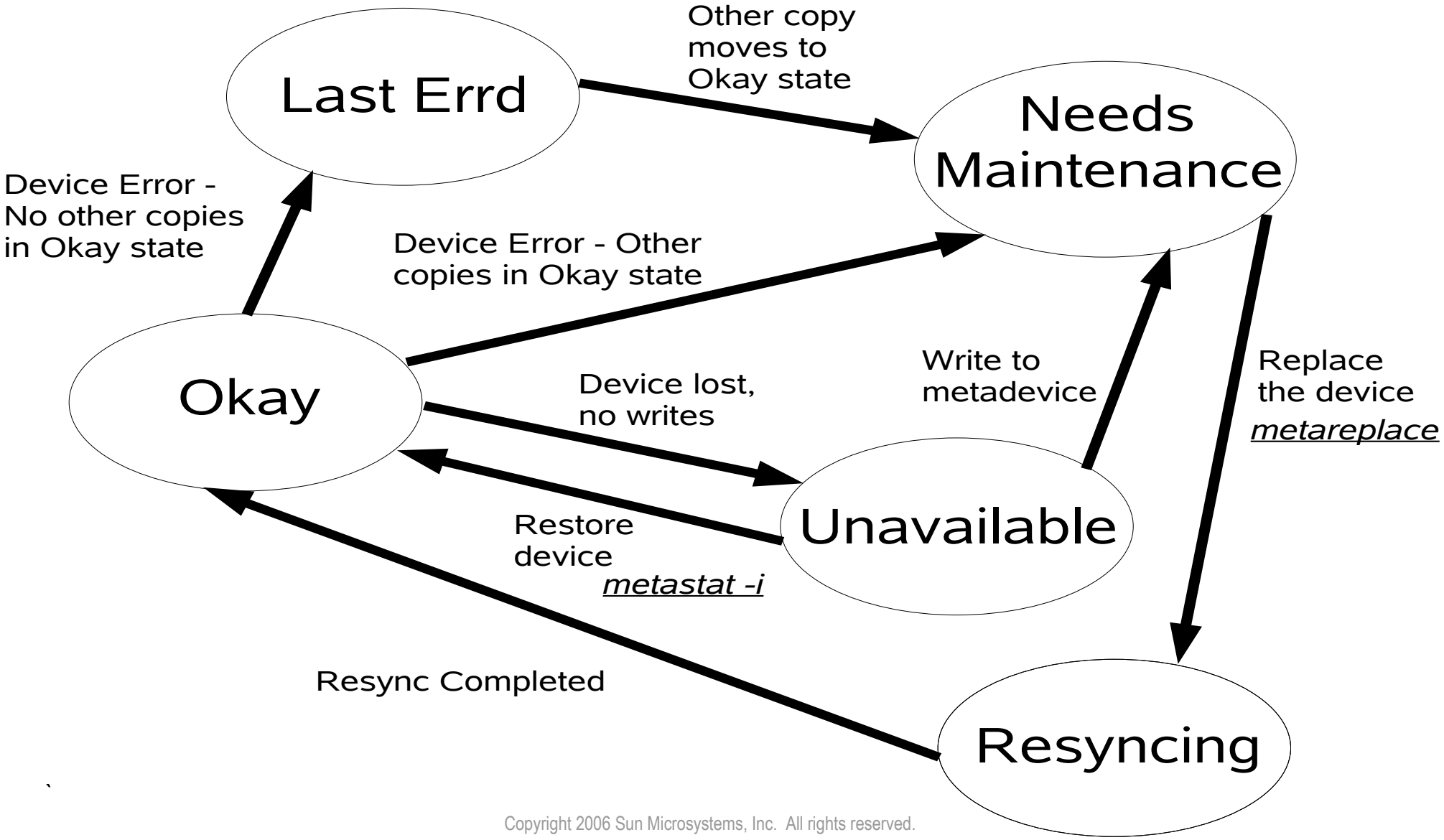


State Model

- Okay
  - > Everything's working fine on the device
- Needs Maintenance
  - > Fault occurred – need to replace the component
- Unavailable
  - > Fault occurred – need to restore the component
  - > No writes have been issued yet
- Last Errd
  - > Last available component, now showing errors

`

State Model



RAID-5 Errors

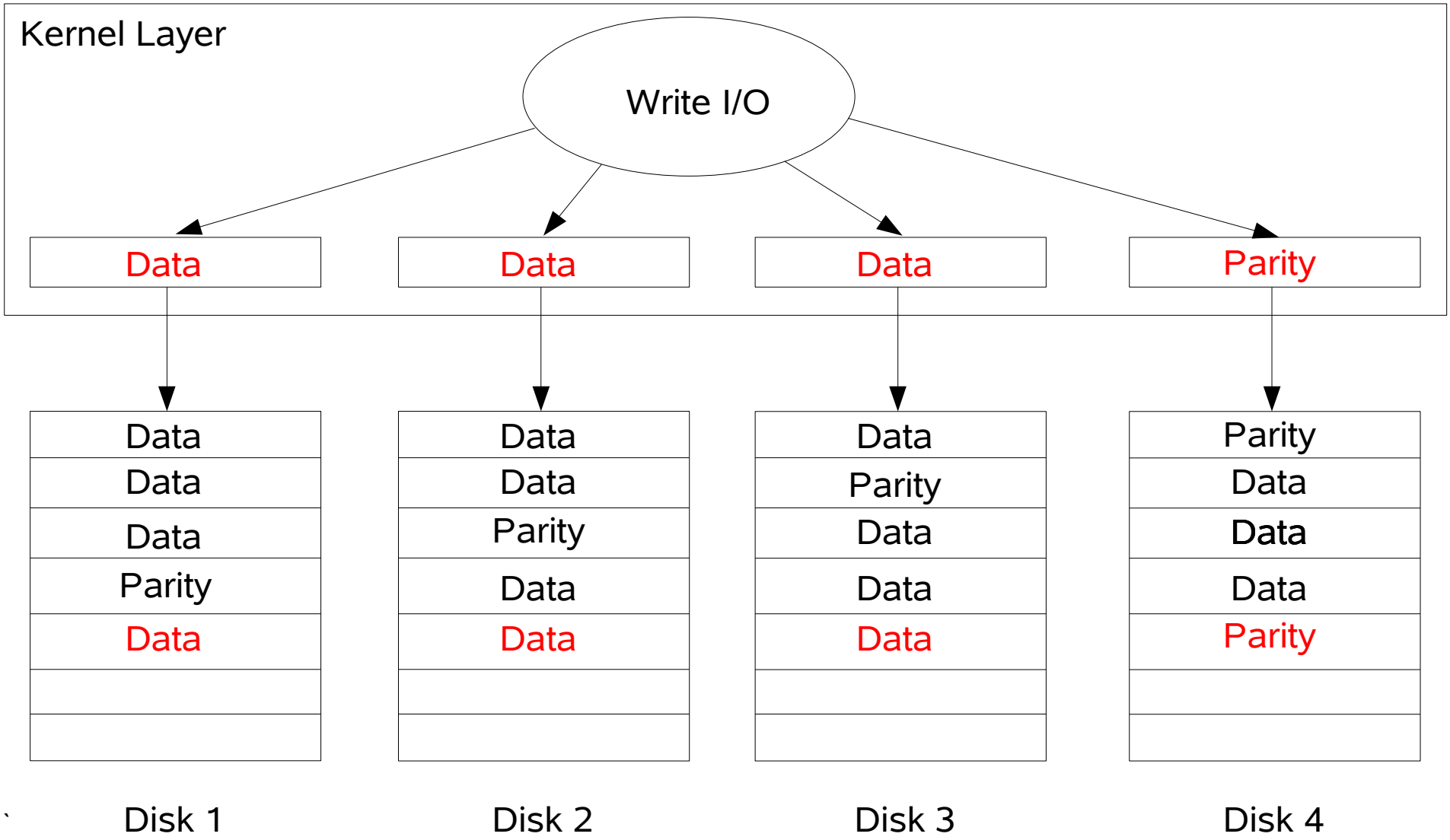
- Error on a component disk
- Data preserved
- Can survive a single-disk failure only
  - > Lose a second device, all data lost
- CPU-intensive to reconstruct data
- Hotspares strongly recommended

RAID-5 Partial-Write Recovery

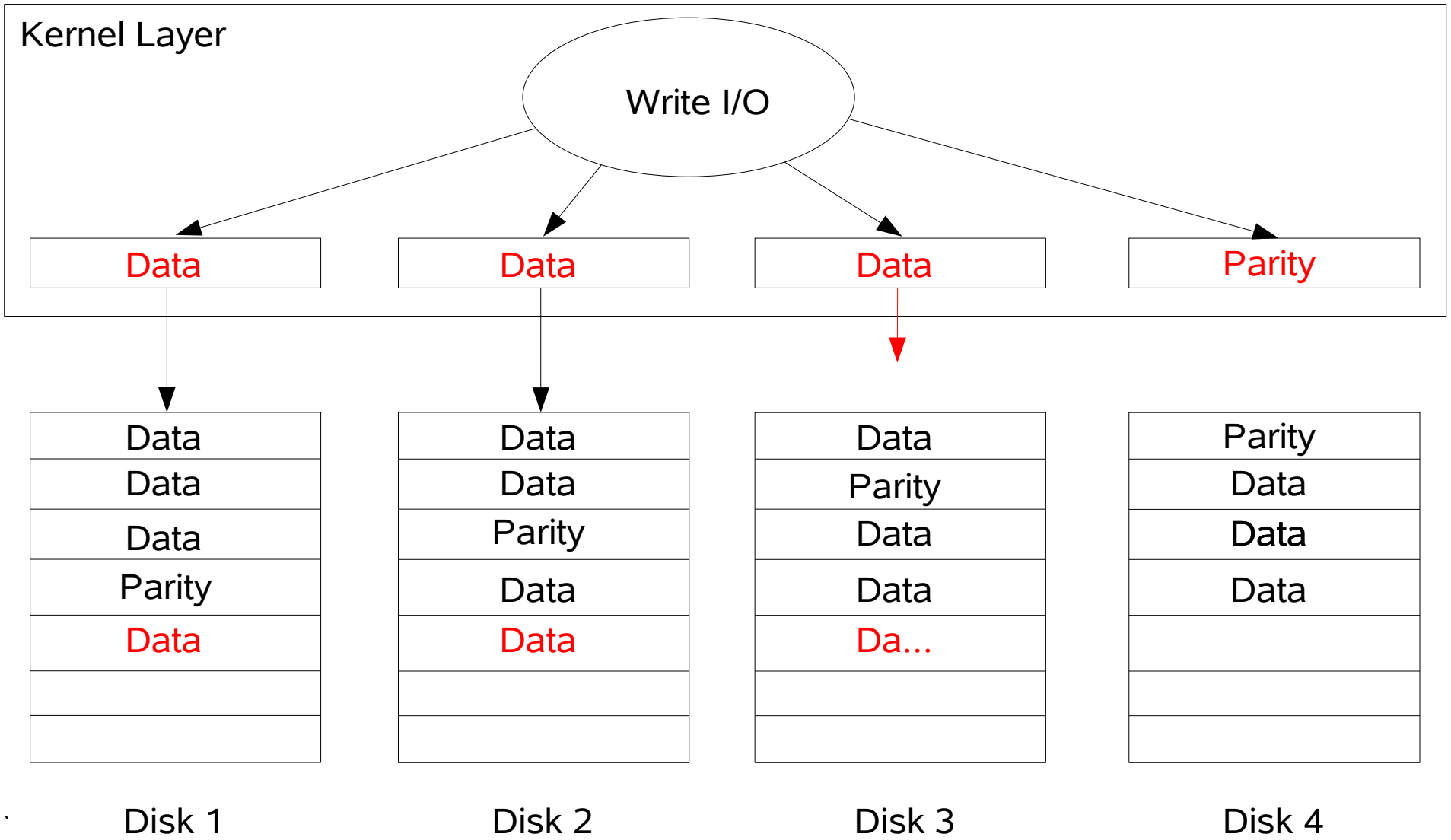
- System crash during RAID-5 write can destroy volume
- Partial-write means not enough data / parity to recover
- Solution – Pre-Write Area

`

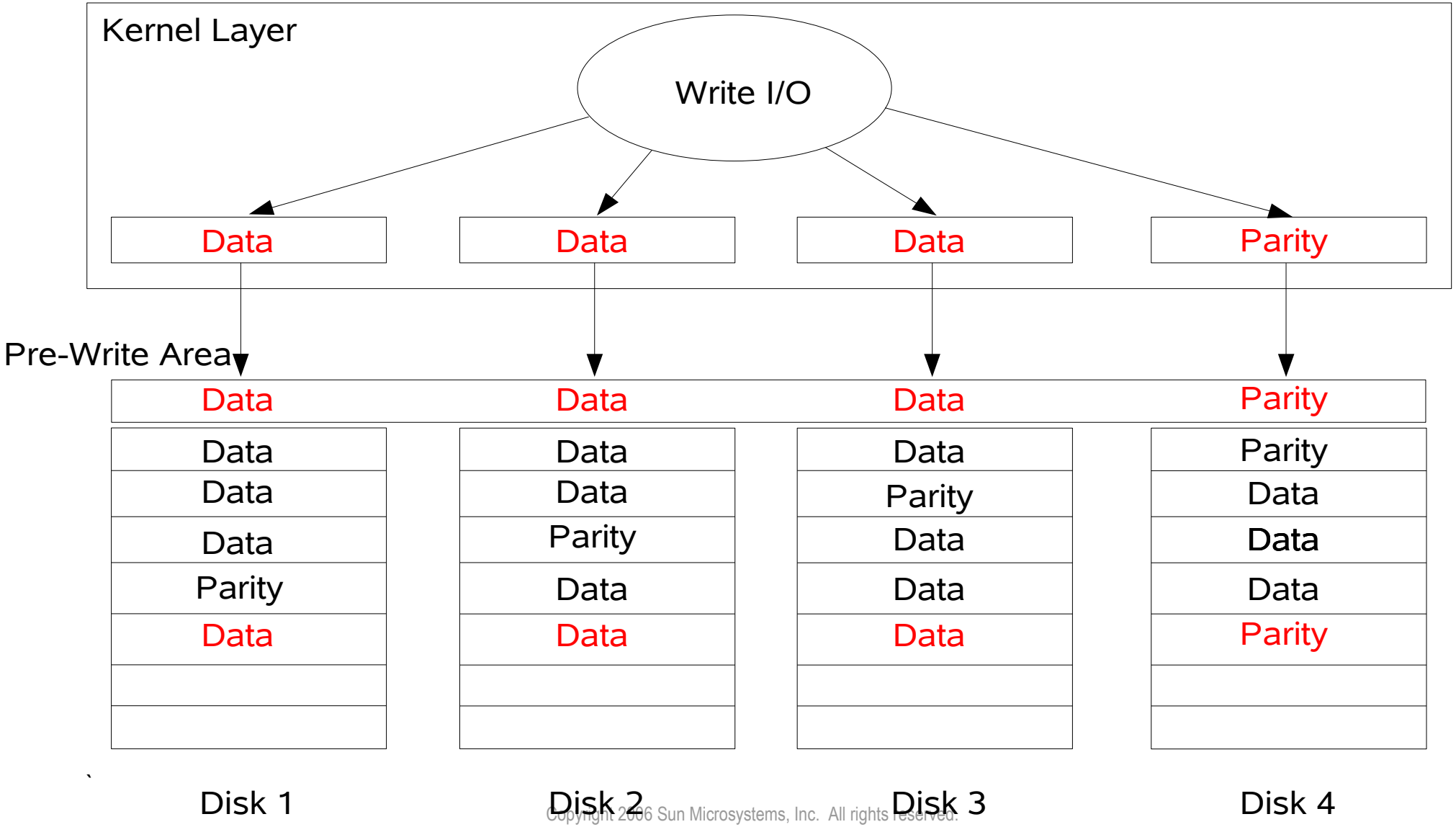
RAID-5 Partial-Write Recovery



RAID-5 Partial-Write Recovery



RAID-5 Partial-Write Recovery



SVM Disk Error Handling