Block Storage Replication with MARS Light

LCA2013 Presentation by Thomas Schöbel-Theuer

Agenda

- □ Differences DRBD vs MARS Light
- □ Operating Principle
- □ Current Status
- □ Appendix: Performance

Multiversion Asynchronous Replicated Storage

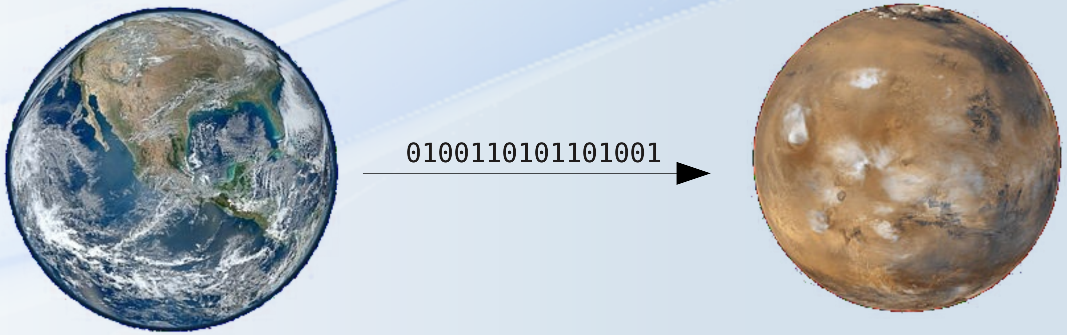


Image source: Wikipedia

Differences DRBD vs MARS Light

DRBD

Application area:

- Distance < 50 km
- synchronously
- Needs reliable network
  "RAID-1 over network"
- Short inconsistencies
  during re-sync
- Low space overhead

currently beta

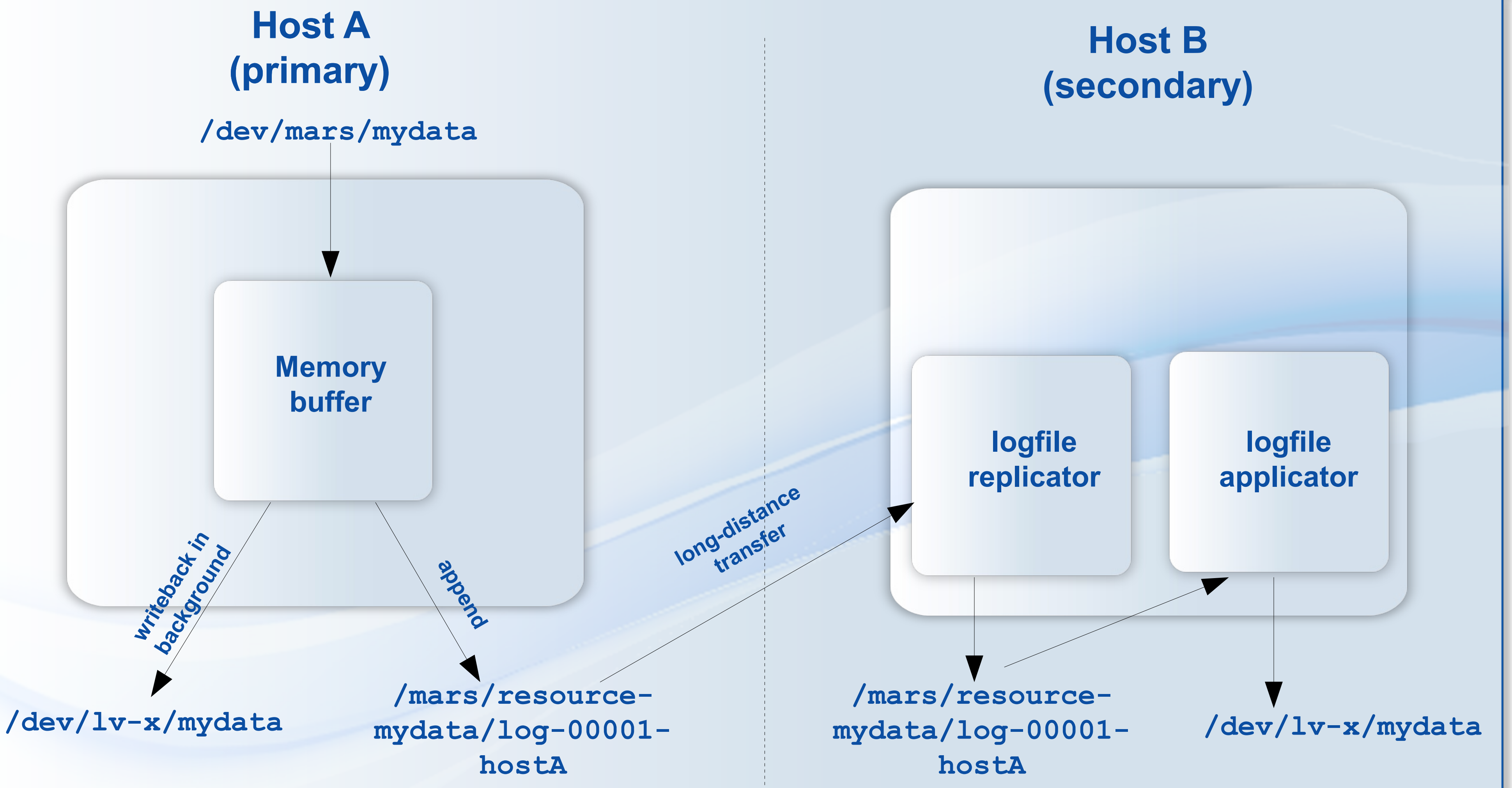
MARS Light

Application area:

- Distances: **any** (>>50 km)
- asynchronously
- Tolerates **unreliable network**
- Anytime consistency
  at cost of actuality (not always up-to-date)
  see Einstein: there exists no coincidence
  in truly Distributed Systems
- Needs >= 100 GB in /mars/
  for transaction logfiles
  dedicated spindle(s) recommended
  RAID with BBU recommended

® 1&1 Internet AG 2012

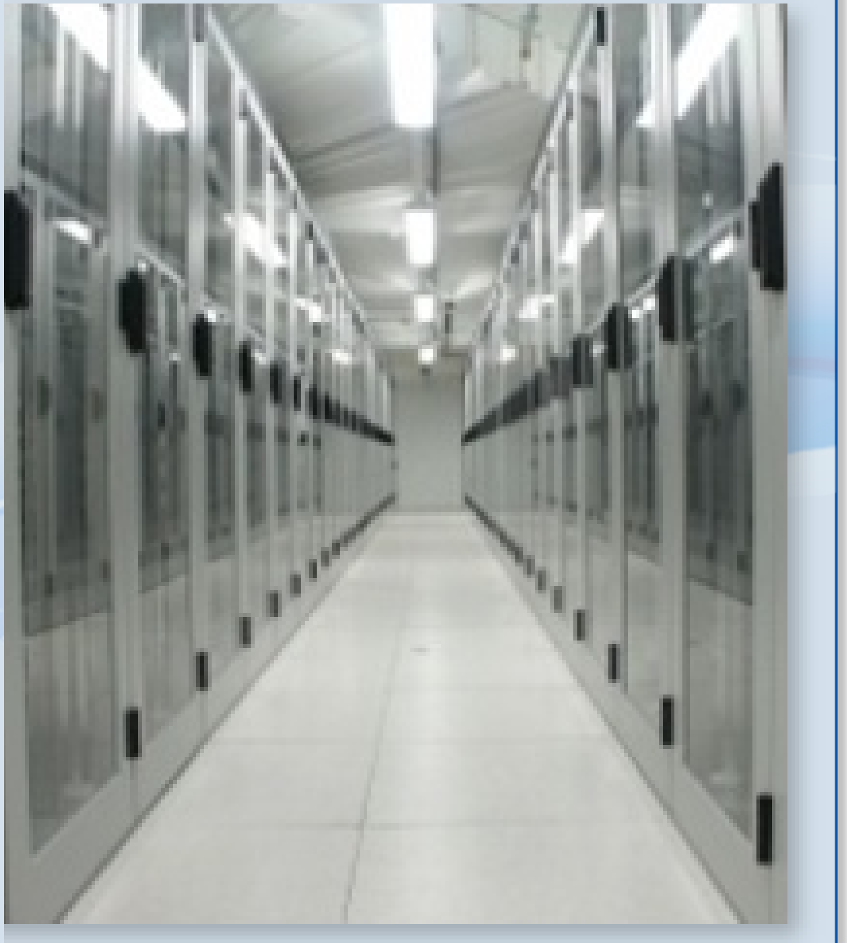
MARS Light: Operating Principle



® 1&1 Internet AG 2012

MARS Light: Current Status

- Beta on `http://github.com/schoebel/mars`
- Linux kernel module under GPL
- Internal pilot system running since 02/2012
  statistics server with highly random IO
- Almost plugin compatible with DRBD
  Example: `marsadm primary mydata`
- Next step: enterprise grade, rollout to >100 servers



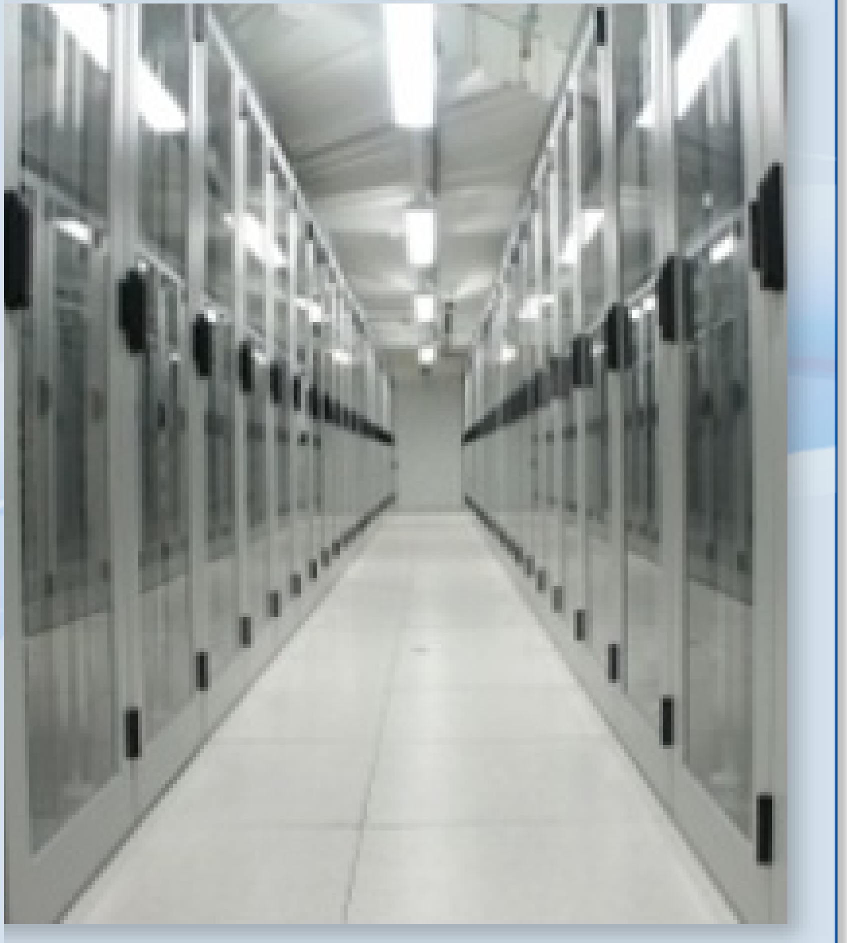
Thanks!

Appendix: MARS Light Performance

- Experimental result: unreplicated mode >50% better performance than RAW device on SATA RAID-6
  - replicated mode: almost no difference (thanks to *sequential* logfiles)

- Preconditions:
  - Load has ~70% random writes
  - Data remains on RAID-6 with BBU
  - /mars/ on RAID-0 (same BBU)
  - ~4 GB RAM for memory buffer

- Internally considered for speedup of *unreplicated* systems (standalone mode)



® 1&1 Internet AG 2012