Reliably Replicating Block Devices
even over Long Distances

LCA2014 Presentation by Thomas Schöbel-Theuer
Agenda

- Use Cases DRBD/proxy vs MARS Light
- Working Model
- Behaviour at Network Bottlenecks
- Current Status / Future Plans
**DRBD (GPL)**

**Application area:**
- Distances: **short** (<50 km)
- Synchronously
- Needs **reliable** network
  - “RAID-1 over network”
  - best with crossover cables
- Short inconsistencies during re-sync
- Under pressure: long or even permanent inconsistencies possible
- Low space overhead

**MARS Light (GPL)**

**Application area:**
- Distances: **any** (>50 km)
- Asynchronously
  - near-synchronous modes in preparation
- Tolerates **unreliable network**
- Anytime consistency
  - no re-sync
- Under pressure: no inconsistency
  - possibly at cost of actuality
- Needs >= 100GB in `/mars/` for transaction logfiles
  - dedicated spindle(s) recommended
  - RAID with BBU recommended
Use Cases DRBD+proxy vs MARS Light

DRBD+proxy
(proprietary)

Application area:
- Distances: any
- Asynchronously
  - Buffering in RAM
- Unreliable network leads to frequent re-syncs
  - RAM buffer gets lost
  - at cost of actuality
- Long inconsistencies during re-sync
- Under pressure: permanent inconsistency possible
- High memory overhead

MARS Light
(GPL)

Application area:
- Distances: any (>>50 km)
- Asynchronously
  - near-synchronous modes in preparation
- Tolerates unreliable network
- Anytime consistency
  - no re-sync
- Under pressure: no inconsistency
  - possibly at cost of actuality
- Needs >= 100GB in /mars/
  for transaction logfiles
  - dedicated spindle(s) recommended
  - RAID with BBU recommended
MARS Working Model

Multiversion Asynchronous Replicated Storage

Datacenter A
(Primary)

/dev/mars/mydata

mars.ko

/dev/lv-x/mydata

/mars/transaction-logfile

Similar to MySQL replication

Datacenter B
(Secondary)

/dev/lv-x/mydata

mars.ko

/mars/transaction-logfile

MARS LCA2014 Presentation by Thomas Schöbel-Theuer
Network Bottlenecks (1) DRBD

- network throughput

DRBD throughput

(potential) incident

automatic disconnect

automatic re-connect

additional throughput needed for re-sync, not possible

wanted application throughput, not possible

decreasing throughput limit

Permanently inconsistent!

mirror inconsistency ...

MARS LCA2014 Presentation by Thomas Schöbel-Theuer
Network Bottlenecks (2) MARS

- network throughput
- time

- MARS
  - application throughput, recorded in transaction log
  - replication network throughput
  - Best possible behaviour at information theoretic limit (better than DRBD)
  - decreasing throughput limit
Network Bottlenecks (3) MARS

- Network throughput
- MARS application throughput
- MARS network throughput
- Flaky throughput limit
- Corresponding DRBD inconsistency

Best possible behaviour (better than DRBD)
Current Status / Future Plans

- Source / docs at
  github.com/schoebel/mars
  or http://mars.technology

- 15 pilot clusters since June 2013

- Rollout project to >250 clusters started

- In preparation / challenges:
  - community revision at LKML planned
  - split into 3 parts:
    - Generic brick framework
    - XIO/AIO personality (1st citizen)
    - MARS Light (1st application)
  - hopefully attractive for other developers!
DRBD+proxy Architectural Challenge

DRBD Host A (primary)

Proxy A’

Proxy B’ (essentially unused)

DRBD Host B (secondary)

bitmap A

huge RAM buffer

data queue path (several GB buffered)

completion path (commit messages)

bitmap B

sector #8 #8 #8 #8 #8 #8

same sector #8 occurs n times in queue

n times

=> need log(n) bits for counter

=> but DRBD bitmap has only 1 bit/sector

=> workarounds exist, but complicated

(e.g. additional dynamic memory)

MARS LCA2014 Presentation by Thomas Schöbel-Theuer
MARS Data Flow Principle

Host A (primary)
/dev/mars/mydata

Temporary Memory Buffer

Transaction Logger

/writeback in background
/append
/dev/lv-x/mydata

/hostA/log-00001

Host B (secondary)

Logfile Replicator
/mars/resource/mydata/log-00001/hostA

Long-distance transfer

Logfile Applicator
/append

/dev/lv-x/mydata

MARS LCA2014 Presentation by Thomas Schöbel-Theuer
Framework Architecture for MARS + future projects

**Framework Application Layer**
MARS Light, MARS Full, etc

**Framework Personalities**
XIO = eXtended IO ≈ AIO

**Generic Brick Layer**
IOP = Instance Oriented Programming
+ AOP = Aspect Oriented Programming

- MARS Light
- MARS Full
- ... future
- Strategy bricks
- other future
- Personalities and their bricks

- XIO bricks
- Generic Bricks
- Generic Objects
- Generic Aspects

External Software, Cluster Managers, etc

Userspace Interface marsadm
Appendix: 1&1 Wide Area Network Infrastructure

- Global external bandwidth > 285 GBit/s
- Peering with biggest internet exchanges on the world
- Own metro networks (DWDM) at the 1&1 datacenter locations
IO Latencies over loaded Metro Network (1) DRBD

Load = ~30,000 IOPS on 50 spindles RAID-6 (7x shared-derived from blkreplay.org)

red = write latency
blue = read latency
IO Latencies over loaded Metro Network (2) MARS

Same load as before, same conditions

red = write latency
blue = read latency