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

Use Cases DRBD vs MARS Light

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
- Aynchronously
  - **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:

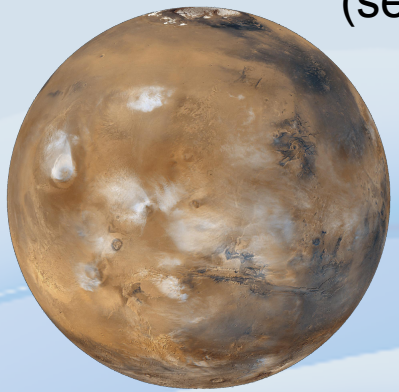
- Distances: **any** (>>50 km)
- Asynchronously
  - near-synchronous modes in preparation
- Tolerates **unreliable network**
- Anytime consistency
  - no re-sync
- Under pressure: no inconsistency
  - possibly at cost of actuality
- Needs >= 100GB in `/mars/` for transaction logfiles
  - dedicated spindle(s)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

MARS Working Model

Multiversion Asynchronous Replicated Storage

Datacenter A
(primary)

Datacenter B
(secondary)



`/dev/mars/mydata`

mars.ko

`/dev/lv-x/mydata`

`/mars/trans-logfile`

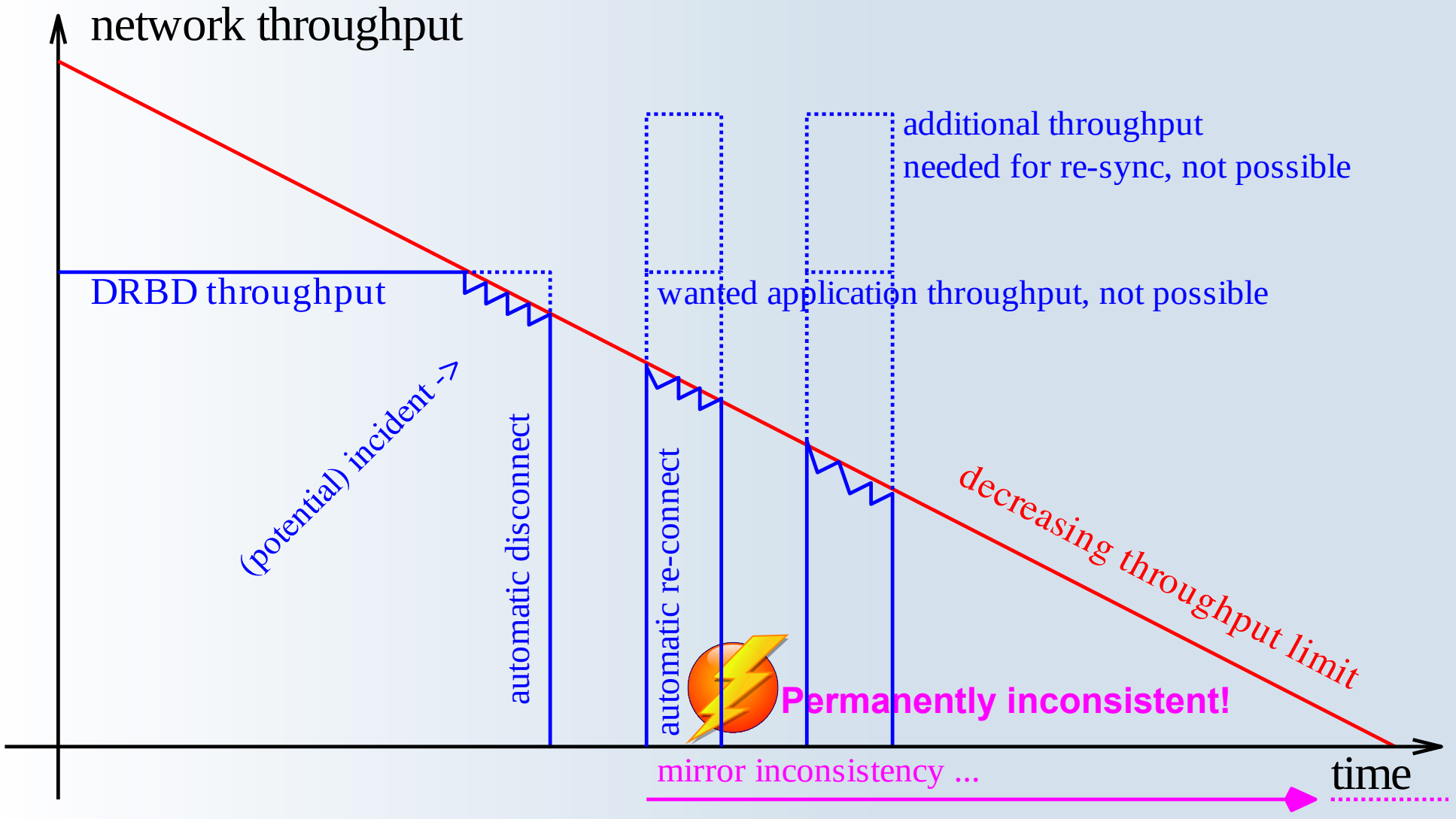
Similar to MySQL replication

mars.ko

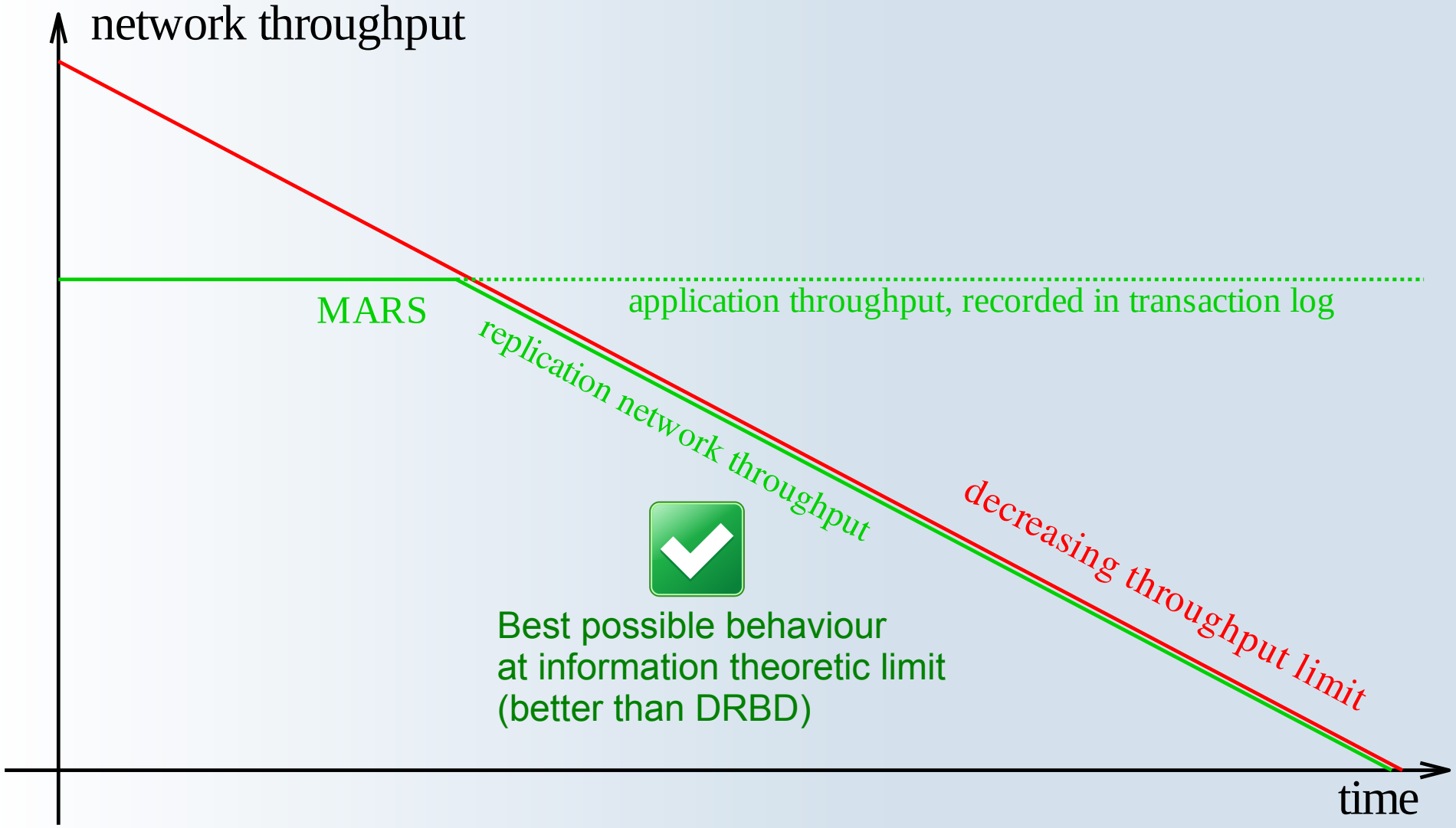
`/mars/trans-logfile`

`/dev/lv-x/mydata`

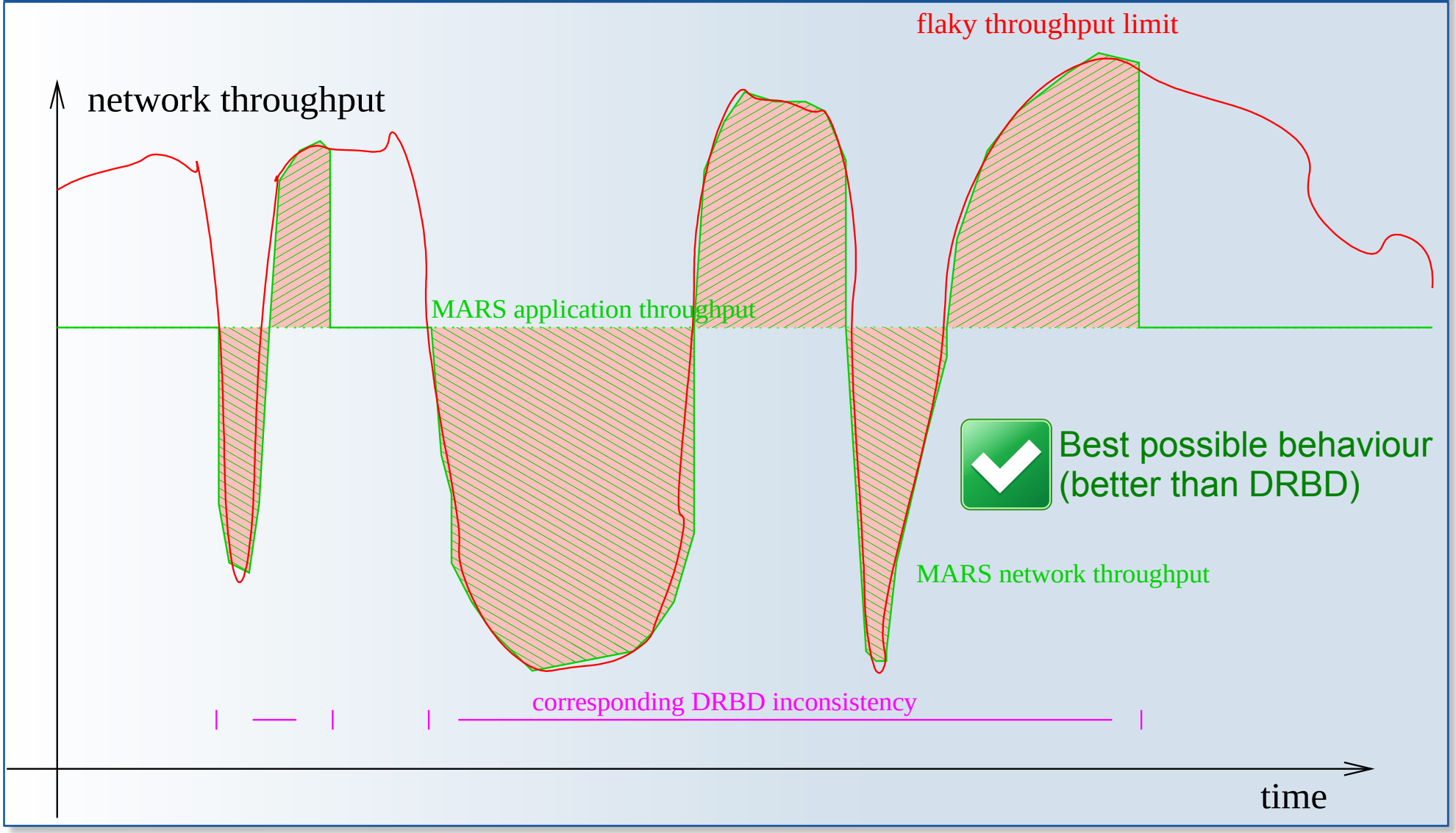
Network Bottlenecks (1) DRBD



Network Bottlenecks (2) MARS



Network Bottlenecks (3) MARS

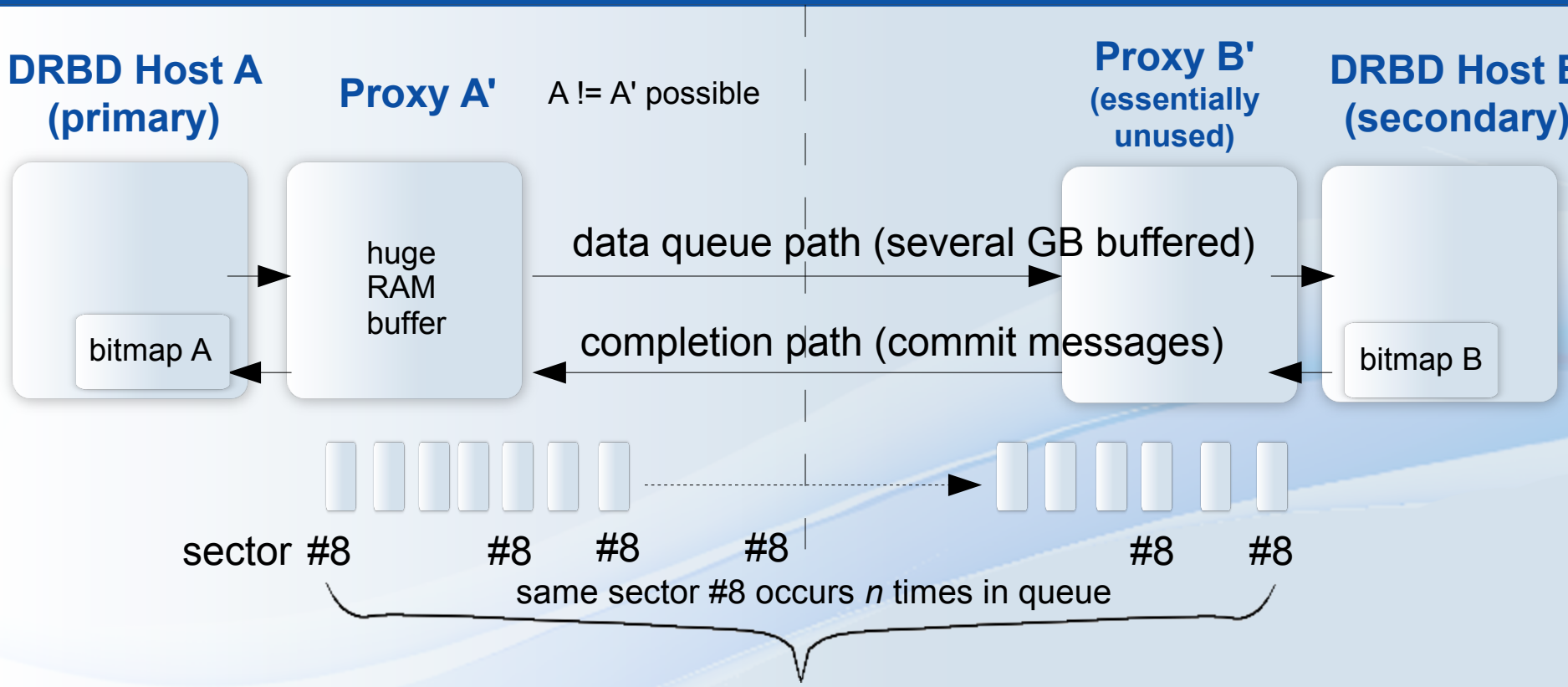


Current Status / Future Plans

- 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!

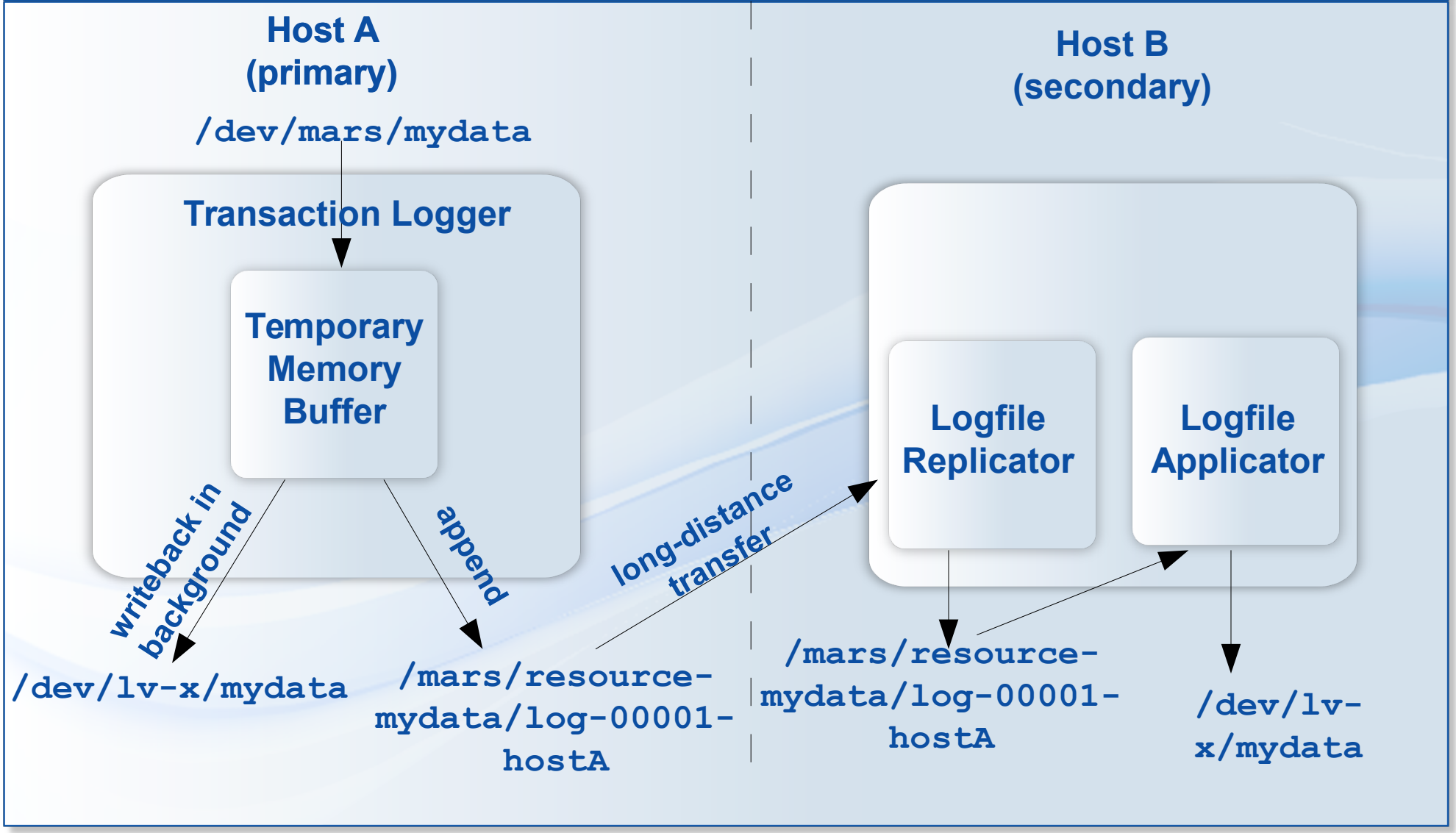
Appendix

DRBD+proxy Architectural Challenge



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 Data Flow Principle



Framework Architecture

for MARS + future projects

External Software, Cluster Managers, etc

Userspace Interface `marsadm`

Framework Application Layer

MARS Light, MARS Full, etc

- MARS Light
- MARS Full
- ...

Framework Personalities

XIO = eXtended IO ≈ AIO

- XIO bricks
- future Strategy bricks
- other future Personalities and their bricks

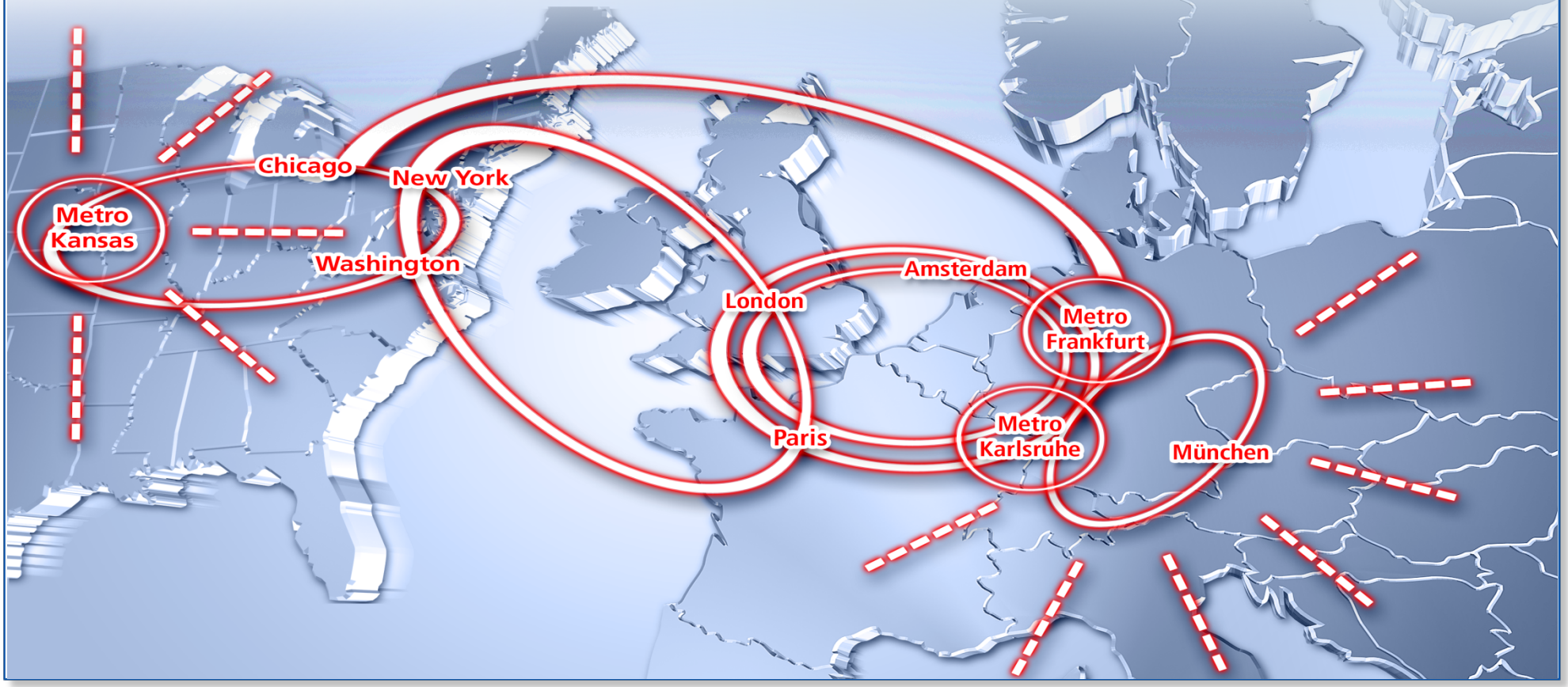
Generic Brick Layer

IOP = Instance Oriented Programming
+ AOP = Aspect Oriented Programming

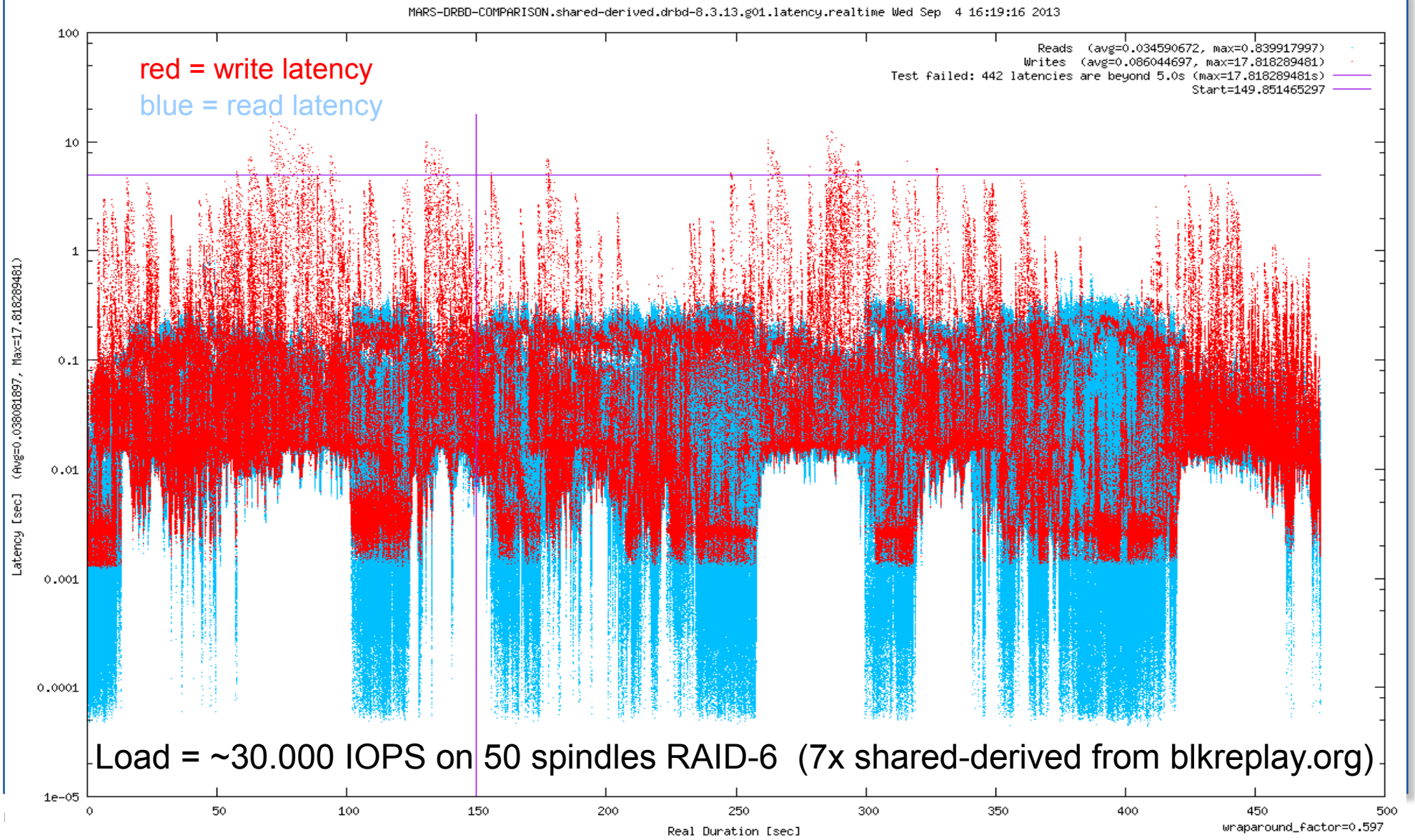
- Generic Bricks
- Generic Objects
- Generic Aspects

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



IO Latencies over loaded Metro Network (2) MARS