Cost-Effective Virtual Petabytes Storage Pools

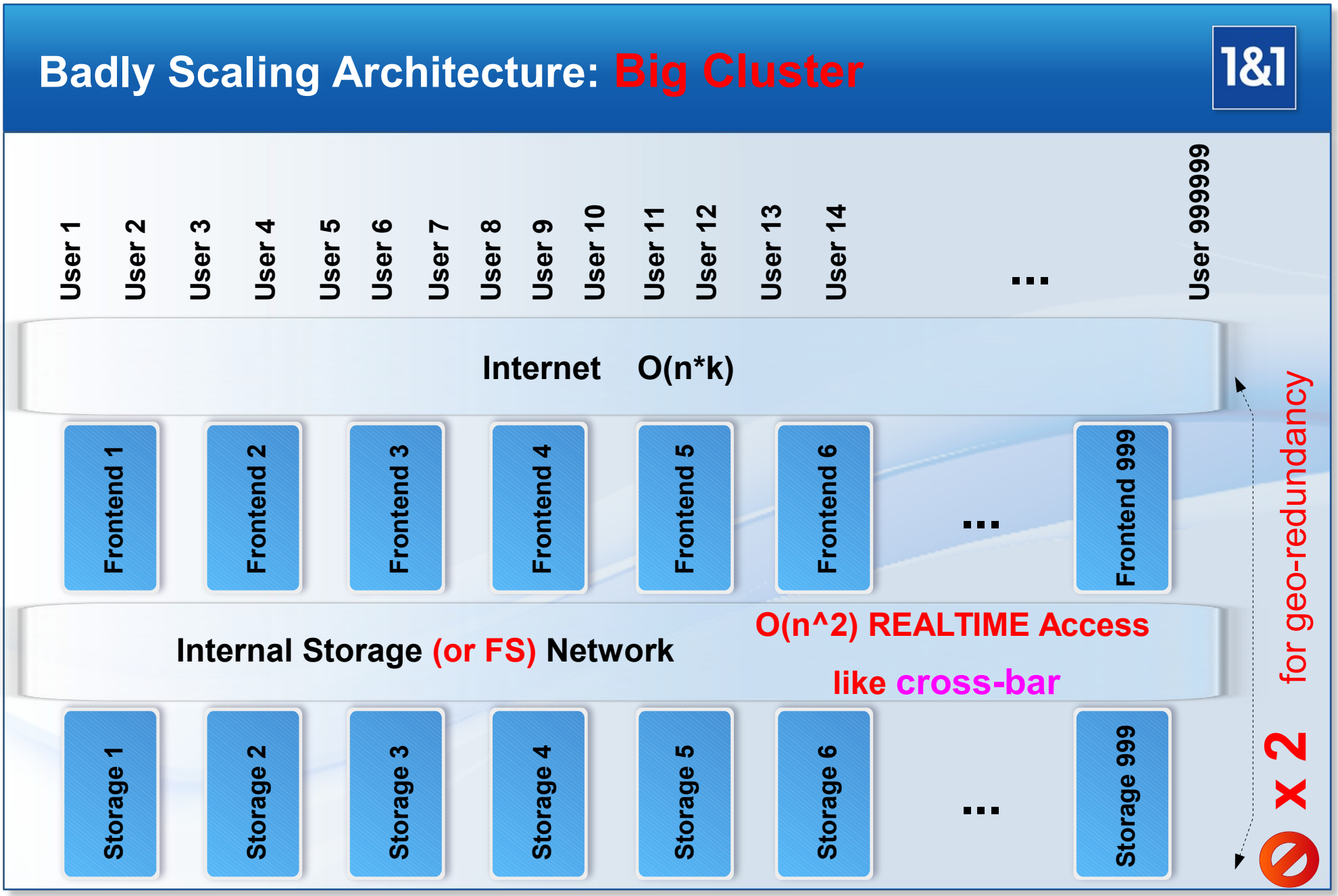
using MARS

GUUG 2017 Presentation by Thomas Schöbel-Theuer

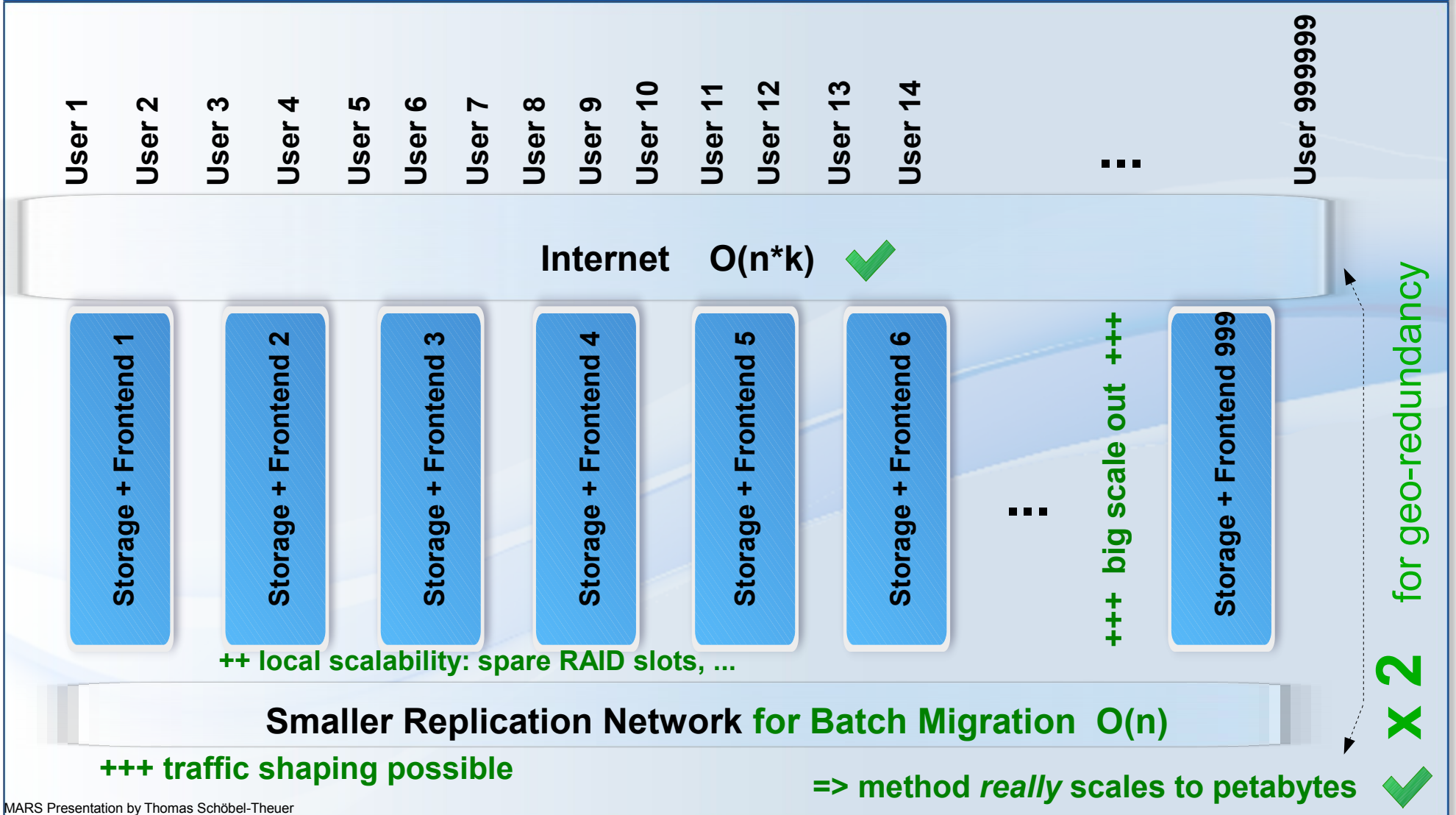
Virtual Petabytes Storage Pools: Agenda

- Scaling Properties of Storage Architectures
- Motivation: Costs
- Flexible MARS Sharding + Cluster-on-Demand
- Load Balancing by Background Data Migration
- Current Status / Future Plans

Badly Scaling Architecture: Big Cluster



Well-Scaling Architecture: Sharding



Costs (non-georedundant Variant)

- Big Cluster:
  - Typically ≈RAID-10 for failure compensation
- Disks: > 200%
- Additional CPU and RAM for storage nodes
- Additional power
- Additional HU

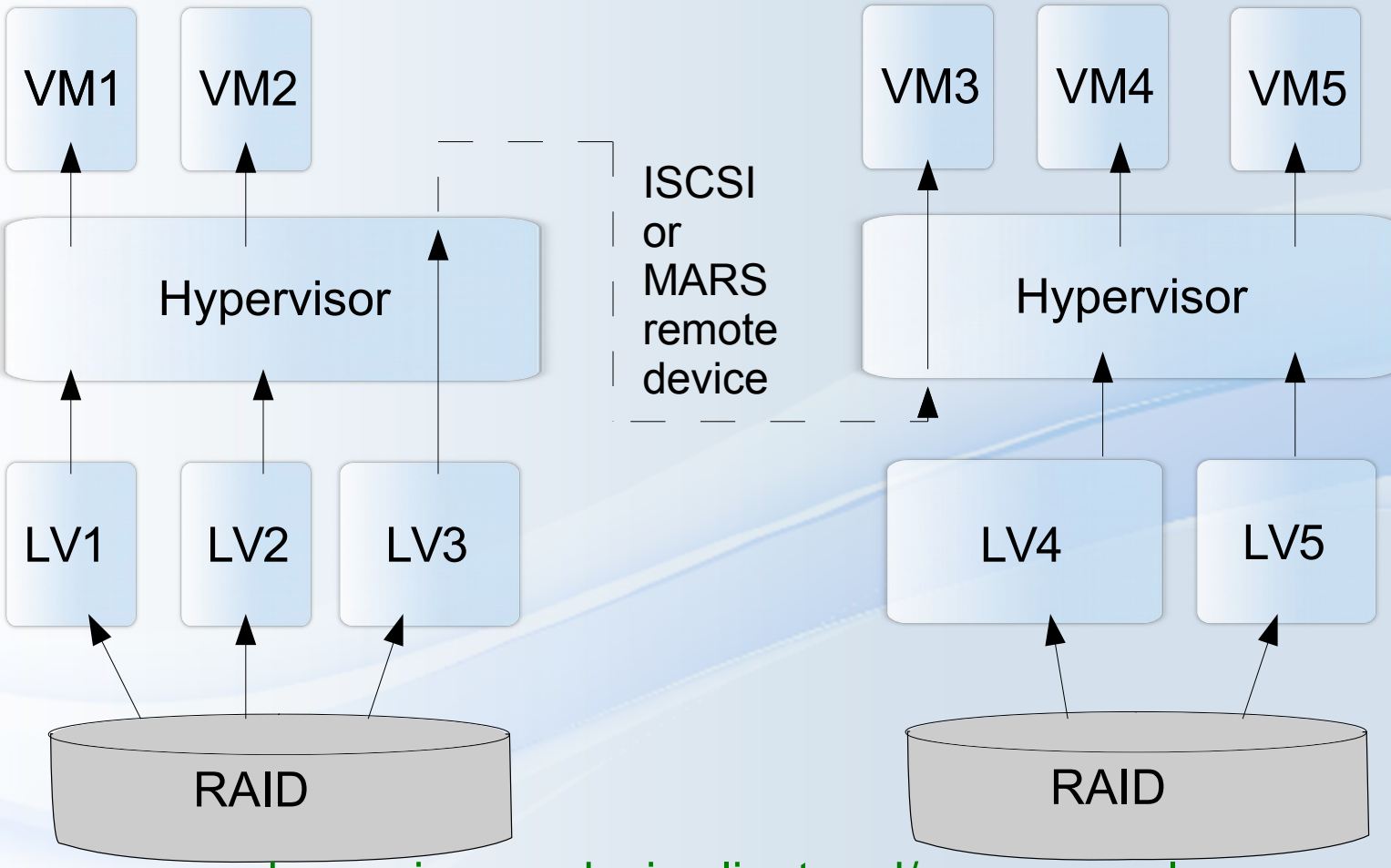
- Sharding:
  - Often local RAID-6 sufficient
- Disks: < 120%
- Hardware RAID controllers with BBU cache on 1 card
- Less power
- Less HU

Costs (georedundant => LONG Distances possible)

- Big Cluster:
  - 2x ≈RAID-10 for failure compensation (smaller does not work in long-lasting DC failure scenarios)
- Disks: > 400%
- Additional CPU and RAM for storage nodes
- Additional power
- Additional HU

- Sharding:
  - 2 x local RAID-6
- Disks: < 240%
- Hardware RAID controllers with BBU
- Less power
- Less HU

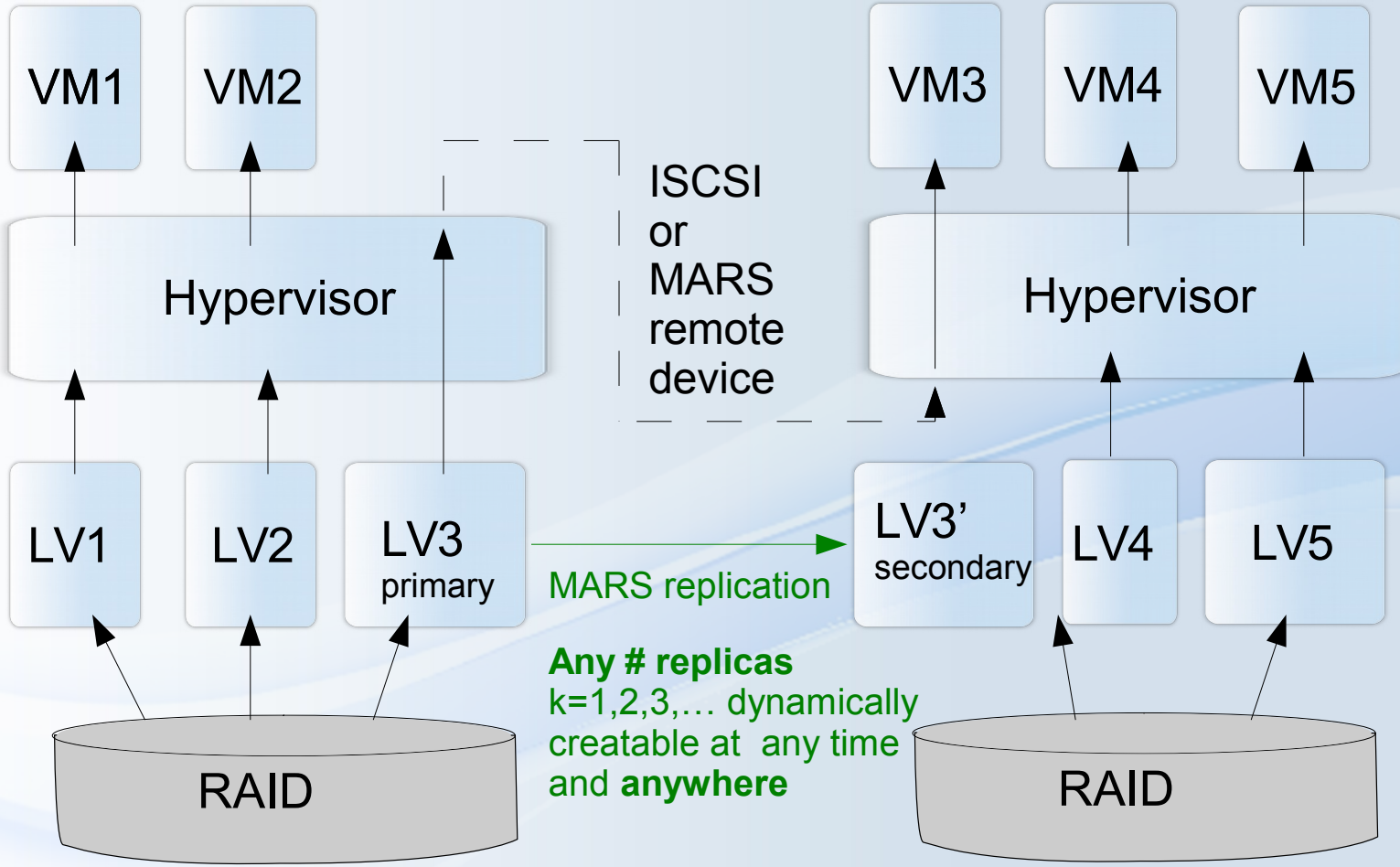
Flexible MARS Sharding + Cluster-on-Demand



any hypervisor works in client and/or server role

and preferably **locally** at the same time

Flexible MARS Background Migration



=> any hypervisor may be source or destination of some LV replicas at the same time

MARS Current Status

- MARS source under GPL + docs:

  `github.com/schoebel/mars`
  `mars-manual.pdf` ~ 100 pages

- mars0.1stable productive on customer data since 02/2014
- Backbone of the 1&1 geo-redundancy feature
- MARS status Feb 2017:

  > 2000 servers (shared hosting + databases)

  > 2x8 petabyte total

  ~ 10 billions of inodes in > 3000 xfs instances

  > 25 millions of operating hours

- New internal Efficiency project
  - Concentrate more LXC containers on 1 hypervisor
  - New public branch mars0.1b with many new features, e.g. mass-scale clustering, socket bundling, remote device, etc
  - mars0.1b currently in ALPHA stage

MARS Future Plans

Automatic load balancing	TBD Separate implementation or libvirt / Openstack plugins … ?
Virtual LVM-like Storage + VM pools	WIP 1&1 clustermanager `cm3` and/or libvirt plugin … ?
Physically sharded pools	Done MARS instead of DRBD

Collaboration sought

=> Opportunities for other OpenSource projects!