RAID is Dead – Long Live CEPH

- **mag. Sergej Rožman**; Abakus plus d.o.o.

- The latest version of this document is available at: http://www.abakus.si/


RAID IS
DEAD
LONG LIVE CEPH !

Abakus
As na disku.

RAID is Dead
Long Live CEPH

mag. Sergej Rožman

sergej.rozman@abakus.si

Make IT
2018

ORACLE Gold Partner

Sava
GOODYEAR
Ljubljana Airport
Mercator
Iskra
MERKUR
Mestna občina Ljubljana
MESTNA OBČINA KOPER
COMUNE CITTA DI CAPODISTRIA
REPUBLIKA SLOVENIJA
MINISTRSTVO ZA FINANCE
BANKA SLOVENIJE
EVROSISTEM
KONTROLA ZRAČNEGA PROMETA SLOVENIJE

Abakus plus d.o.o.

ORACLE Gold Partner

History

- from 1992, 20-30 employees

Applications:

- DejaVu - High Performance Architecture for Virtual Databases
- ARBITER – the ultimate tool in audit trailing
- APPM – Abakus Plus Performance Monitoring Tool

Services:

- DBA, OS administration , programming (MediaWiki, Oracle)
- networks (services, VPN, QoS, security)
- open source

Infrastructure:

- servers, **SAN storage**, firewalls, backup servers

Skills & Experience:

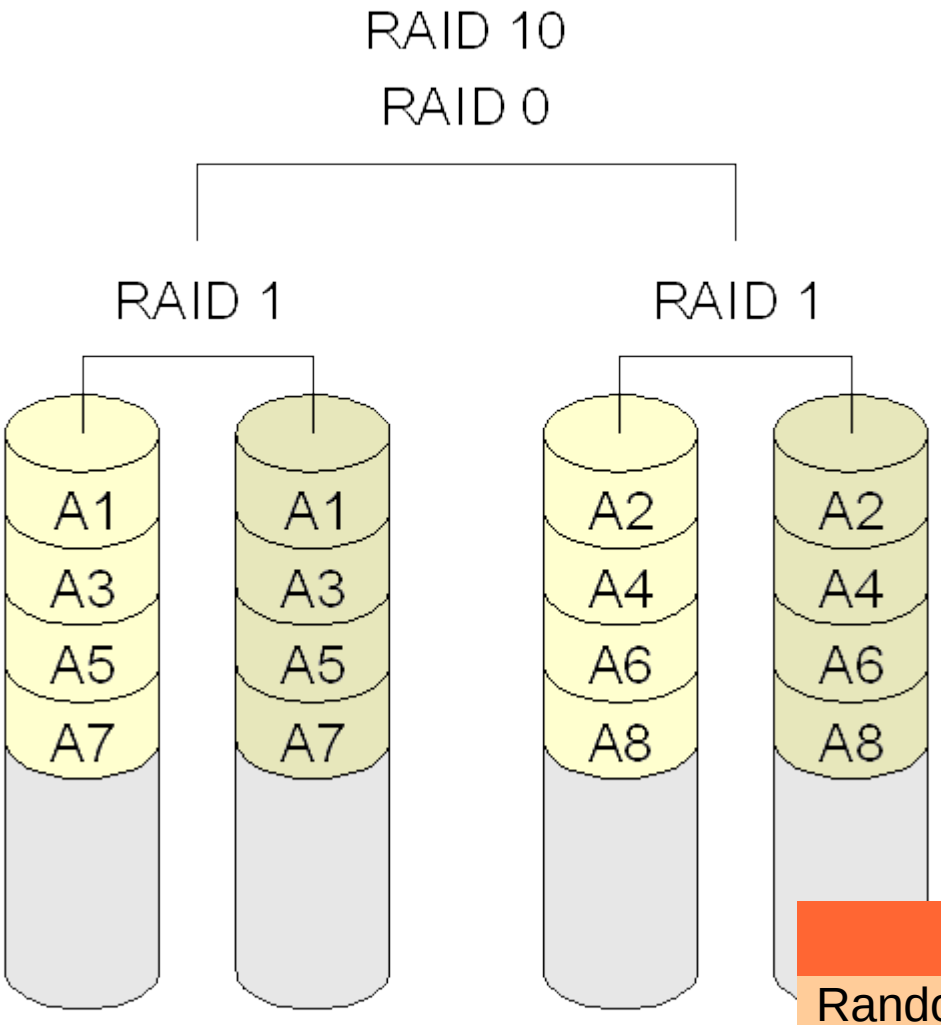
- from 1995 GNU/Linux ***(>20 years of experience !)***
- Oracle on GNU/Linux: since RDBMS 7.1.5 & Forms 3.0 ***(before Oracle !)***
- **~30 years of experience with High-Availability !**

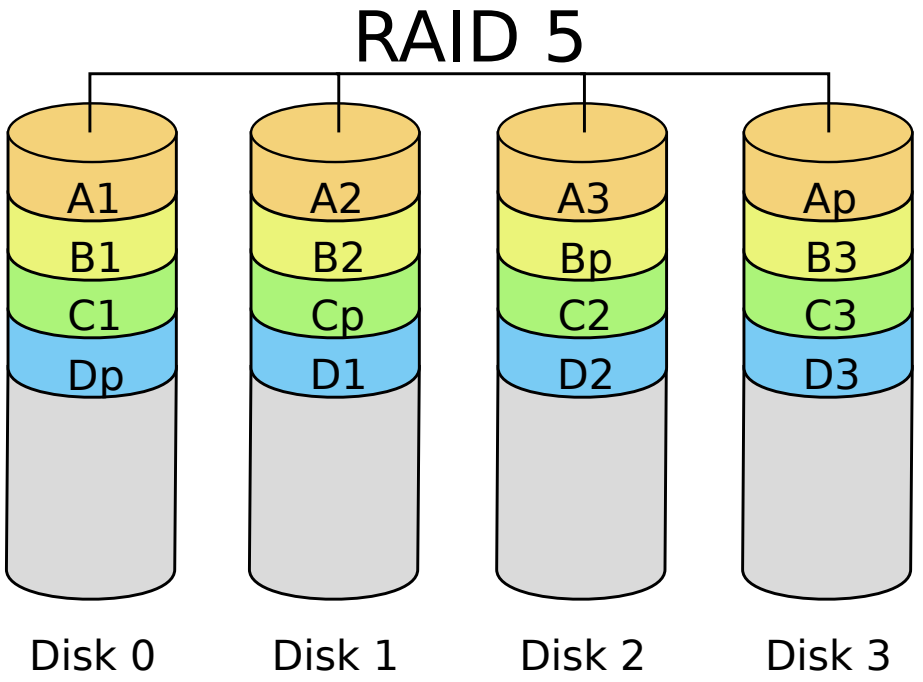
Mestna občina Ljubljana
MESTNA OBČINA KOPER
COMUNE CITTÀ DI CAPODISTRIA
Iskra
Ljubljana Airport
KONTROLA ZRAČNEGA PROMETA SLOVENIJE
NOVA BANKA
GOODYEAR

RAID (Redundant Array of Independent Disks)

- University of California, Berkeley in 1987
  (mirroring long before)
- brilliant solution; the most valuable resource is DATA

RAID Levels



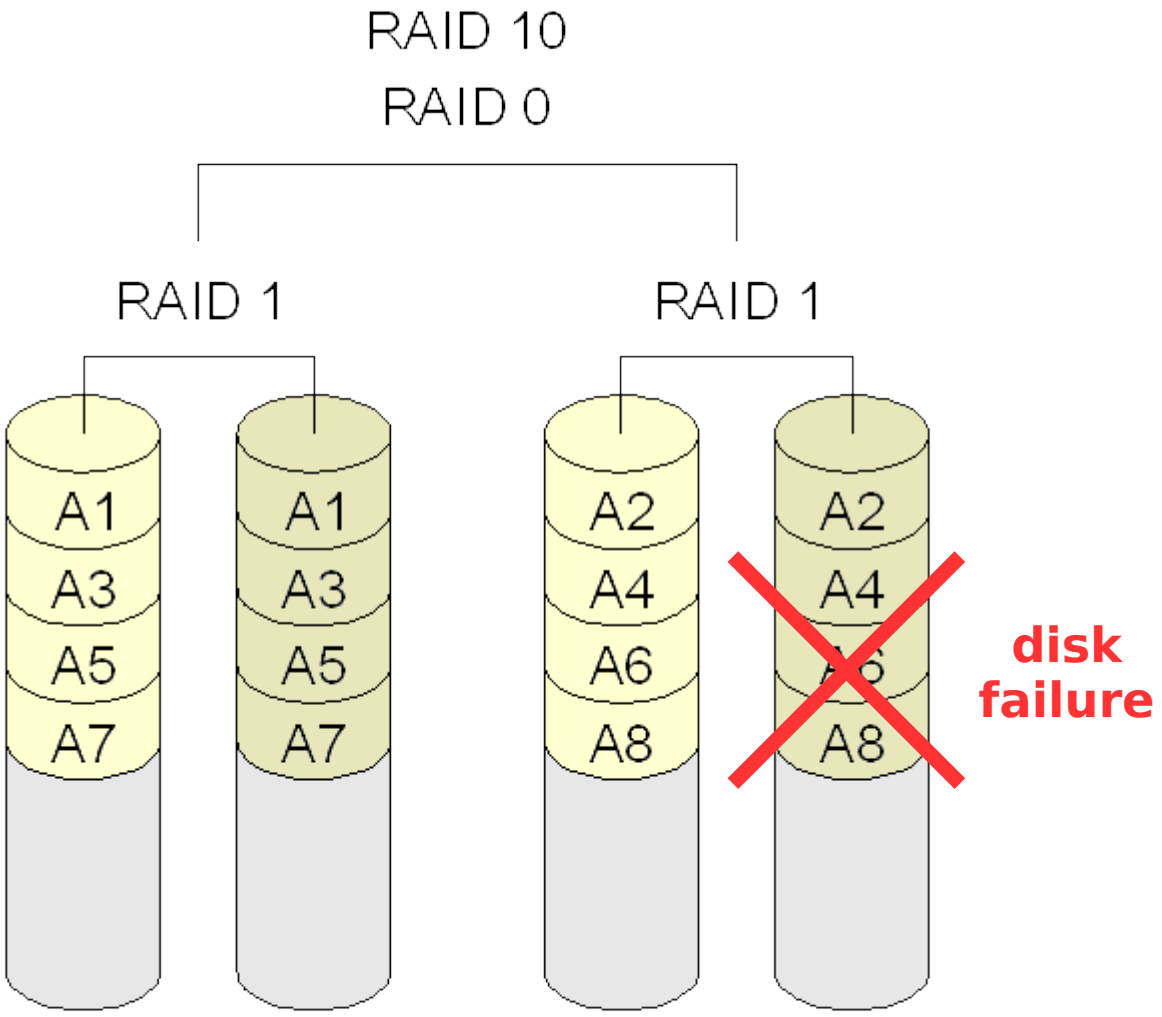


RAID	IOPS
Random Writes RAID10	14.399
Random Writes RAID5	2.703
Random Writes RAID6	1.942

source: wikipedia & HP

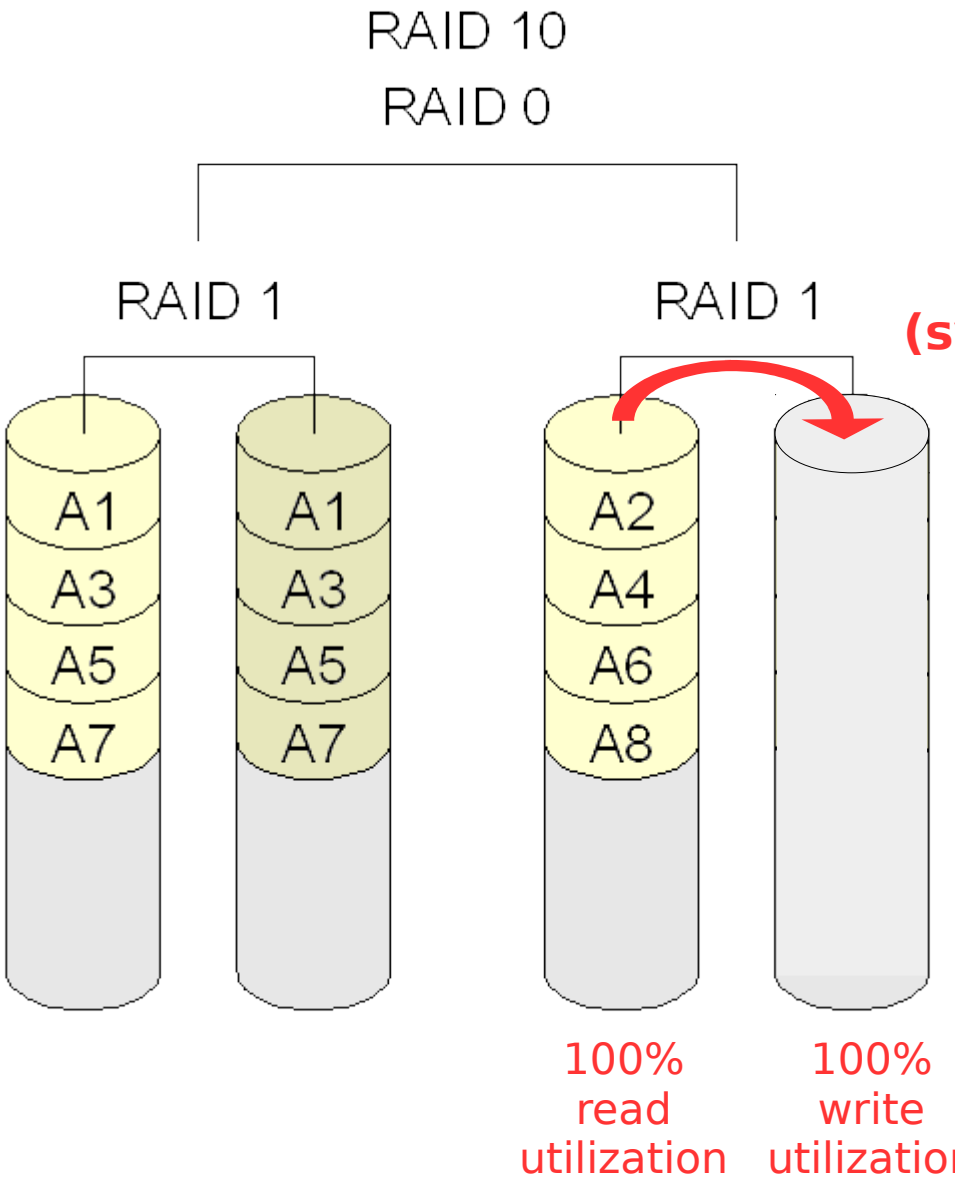
Why RAID has become insufficient?

1. RAID Disk Failure and Recovery



Abakus
As na disku.

RAID Recovery



Synchronization

- high utilization impacts the performance of a whole array
- using TB drives it can take a whole day or more to finish

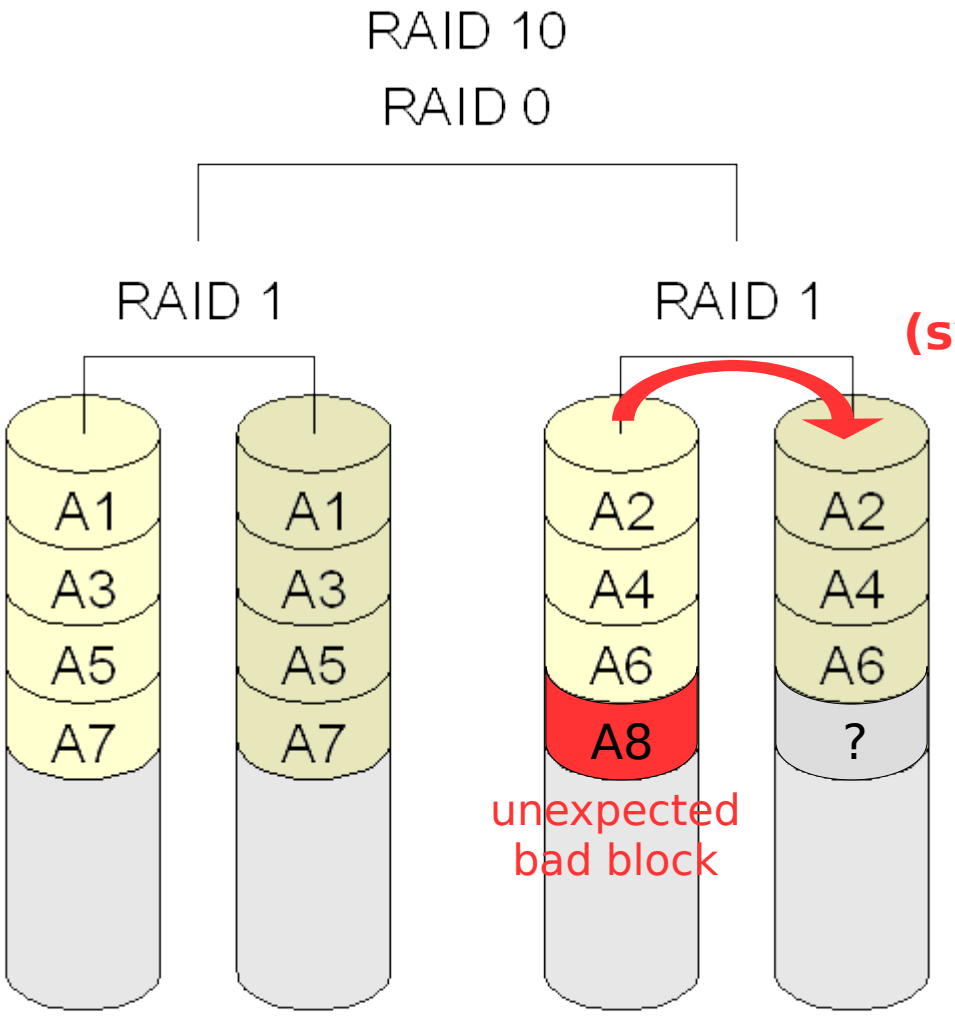
Abakus
As na disku.

RAID Fails at Scale

- bit rate error (BRE) of typical SATA drive is 10^{14}
- when reading 10 terabytes, the probability of an unreadable bit is likely (56%)
- when reading 100 terabytes, it is nearly certain (99.97%).

Reference:
CLEVERSAFE WHITE PAPER:
Why RAID is Dead for Big Data Storage
(http://docs.media.bitpipe.com/io_10x/io_103356/item_496432/Cleversafe%20-%20Why%20RAID%20is%20Dead%20for%20Big%20Data%20Storage.pdf)

RAID Recovery Failure



recovery (synchronization)

Unsuccessful synchronization

- whole array can fail!

Abakus
As na disku.

2. Limited and Slow RAID Reconfiguration

- OLRM (Online RAID Level Migration) and
  OCE (Online Capacity **Expansion**)

Examples

1.

- Expansion from 4-member 2TB RAID10 to 6-member 2TB RAID10 on RocketRAID 2740 **takes more than 300 hours**.

Abaku
As na disku.

Slow RAID Reconfiguration

https://serverfault.com/questions/297072/how-long-should-a-raid-reconfiguration-take-adaptec-6805

2.

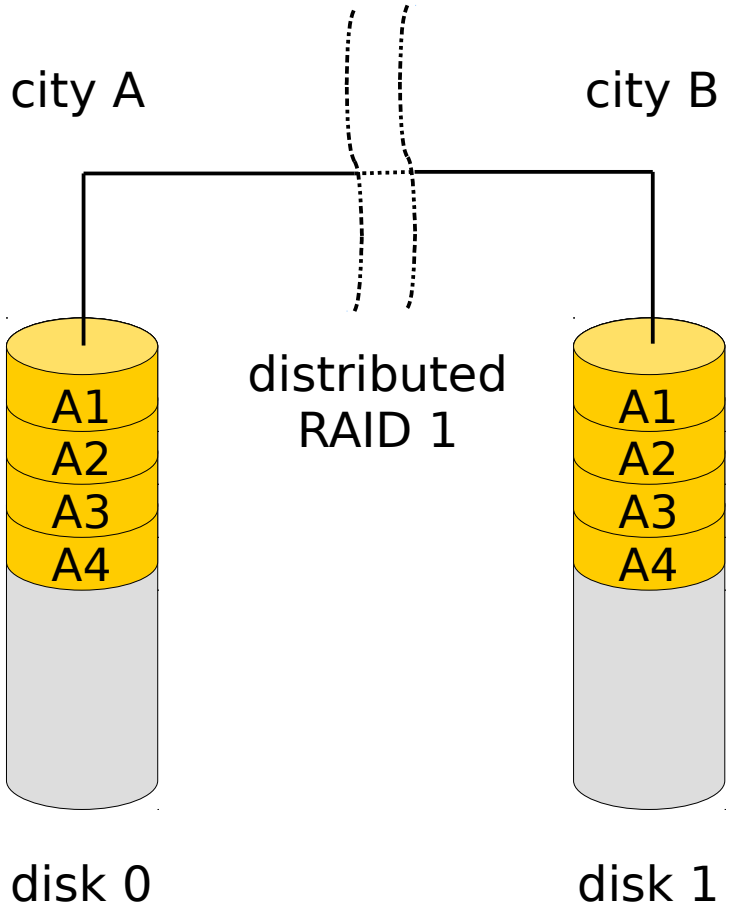
I created a RAID6 array comprising four disks, two from each connector. Then I tried expanding the array to the remainder of the disks.

My problem is that raid reconfiguration never seems to progress. **After more than 24 hours the Adaptec Storage Manager still shows that it's at 0% completion.**

3.

I have an Adaptec 5405Z with 20x2TB 7200 RPM drives on a SAS Backplane. I attempted to do a reconfig on it to go from 8 drives to 20 drives. We use it for Security Video Storage. Since the box was essentially brand new, I figured why not see how long it would take with ~2TB of data on the array.
After about a week and it only getting to 10%, I gave up.

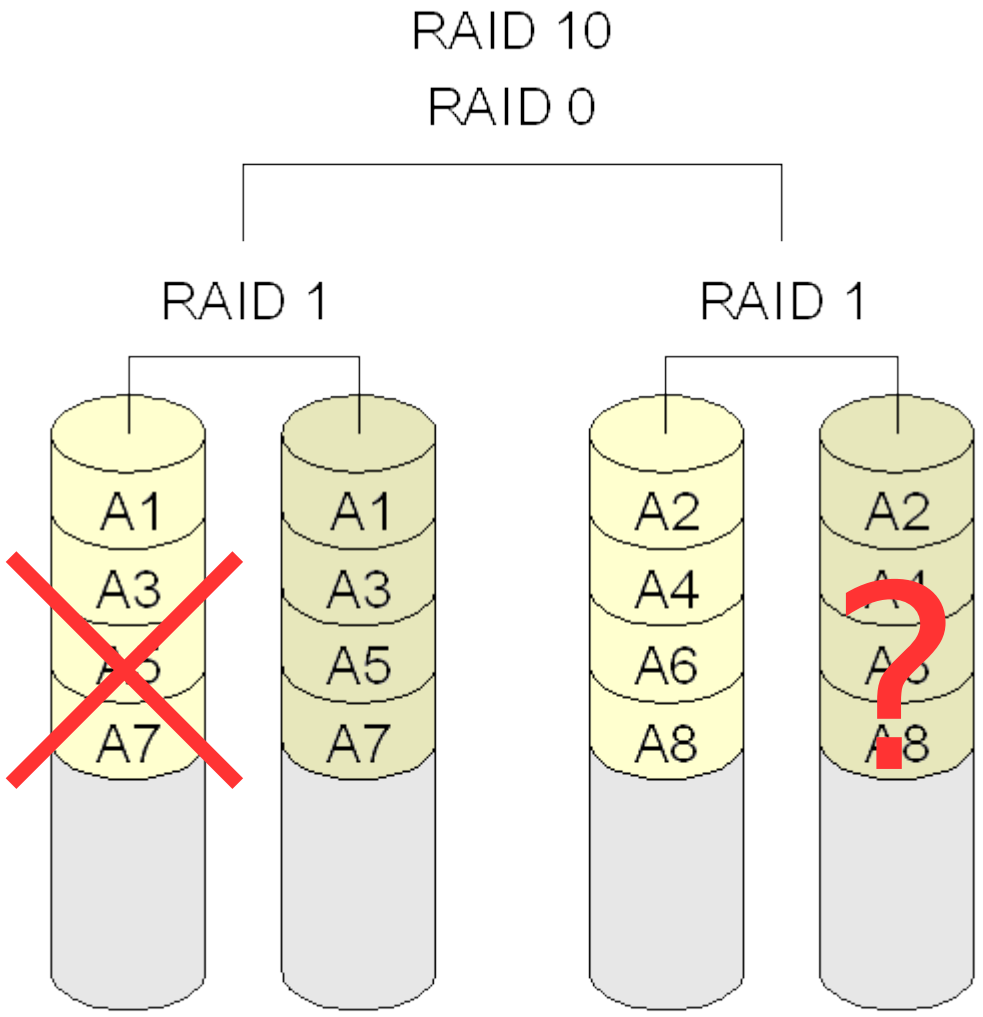
3. Distributed RAID Arrays



- almost nonexistent (especially RAID5 & 6)
- software-based RAID only
  - DRBD
  - odd solutions software-based mirror of iSCSI disks

Abakus
As na disku.

4. Fault Detection and Isolation



wrong disk failure

problematic disk

- limited or no control over single disk drive member
- isolation problem: problematic but not yet failed disk can hurt whole RAID array or even freeze the controller

Abakus
As na disku.

CEPH – The Future of Storage™

- since 2006
- 2012: the first major »stable« release – Argonaut

What is CEPH?

- »infinitely« scalable distributed storage cluster
- runs on commodity hardware !?
- is free

What is a commodity hardware?

Commodity hardware is not an abandoned, deserted, dropped, discarded, legacy hardware from a junkyard!

CEPH requirements:

- 1 GHz CPU core for every served physical disk drive
- 1 GB of RAM for every 1 TB of storage
- lots of disks (SSDs recomended)
- no RAID controller

CEPH Contributors

CANONICAL

CERN

CISCO

FUJITSU

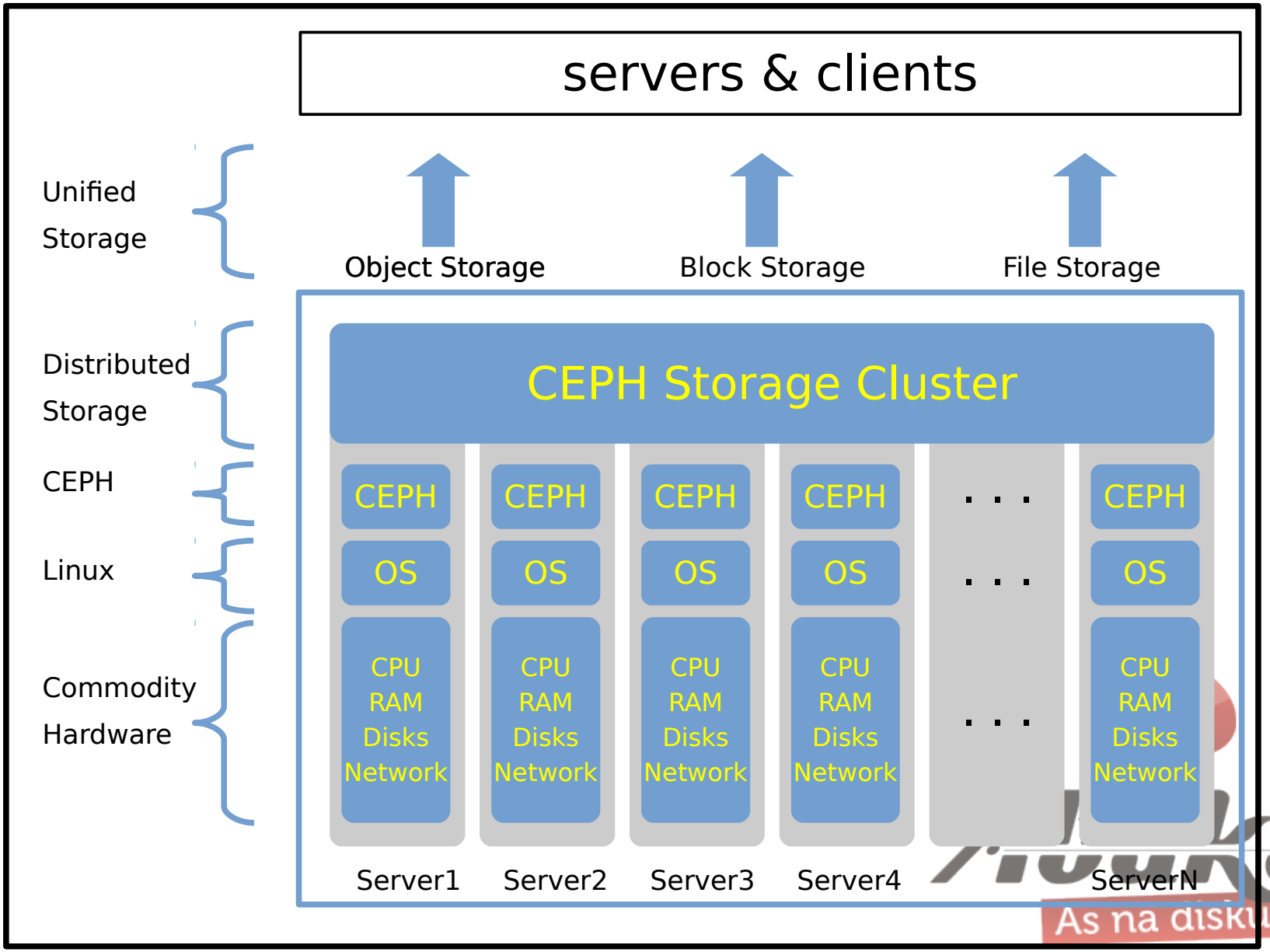
intel

redhat

SanDisk

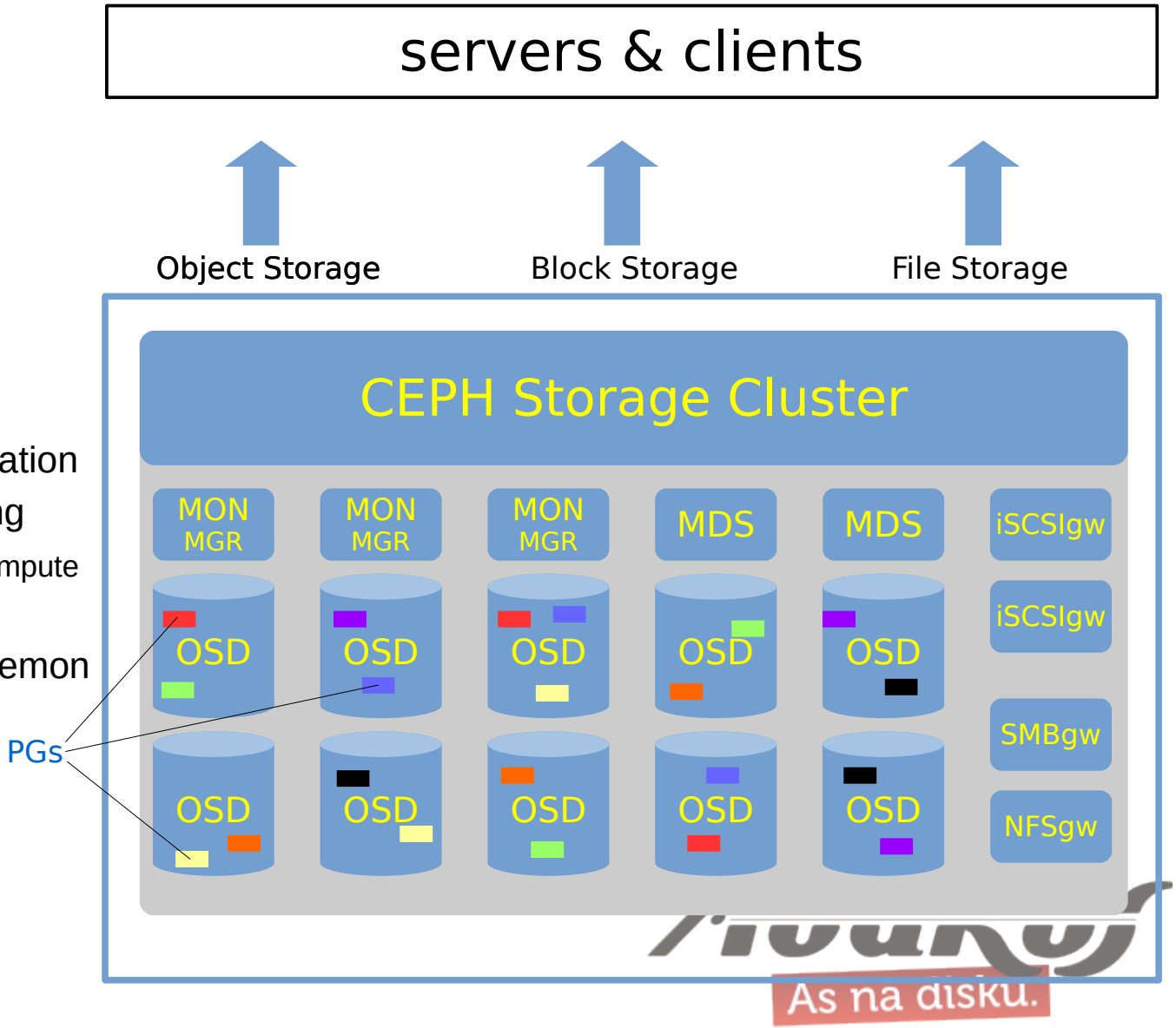
SUSE

CEPH From the Outside

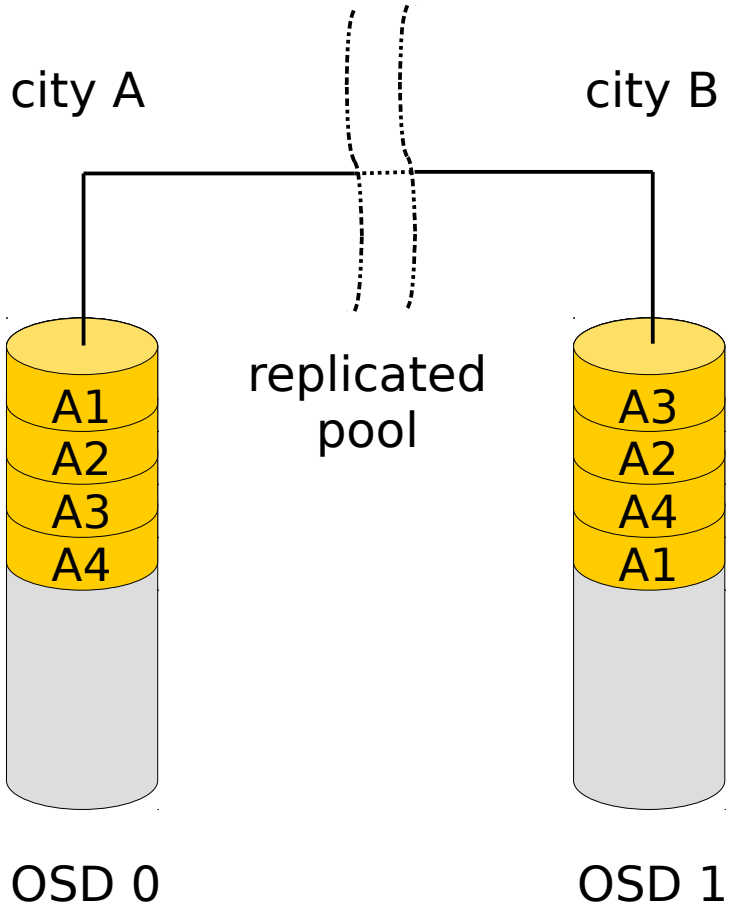


CEPH From the Inside

- OSD - Object Storage Device/Daemon
- PGs – Placement Groups
- CRUSH – Controlled Replication Under Scalable Hashing
  (algorithm CEPH uses to compute PG locations)
- MDS – Metadata Server Daemon



Distributed CEPH Array



CEPH embedded replication types (failure-domains)

- type 0 osd
- type 1 host **<- default**
- type 2 chassis
- type 3 rack
- type 4 row
- type 5 pdu
- type 6 pod
- type 7 room
- type 8 datacenter
- type 9 region
- type 10 root

OSD – Object Storage Device/Daemon

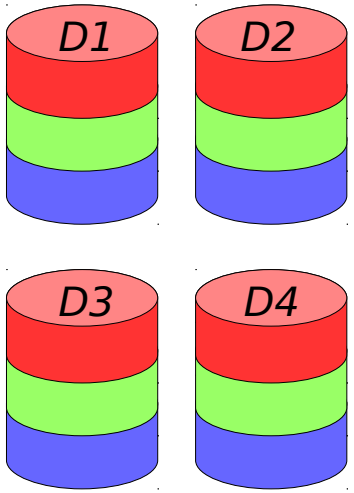
CEPH Top 10 Features

1. no single point of failure
2. »infinite« scalability
3. self managing & self healing
4. high availability authentication
5. thin provisioning
6. snapshots & clones
7. copy on write & copy on read
8. replication (default 3/2) & erasure coding
9. storage tiering
10. remote replication to disaster site

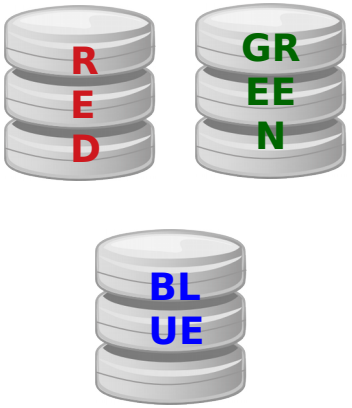
CEPH Usage – Oracle DB Server

DB server 1

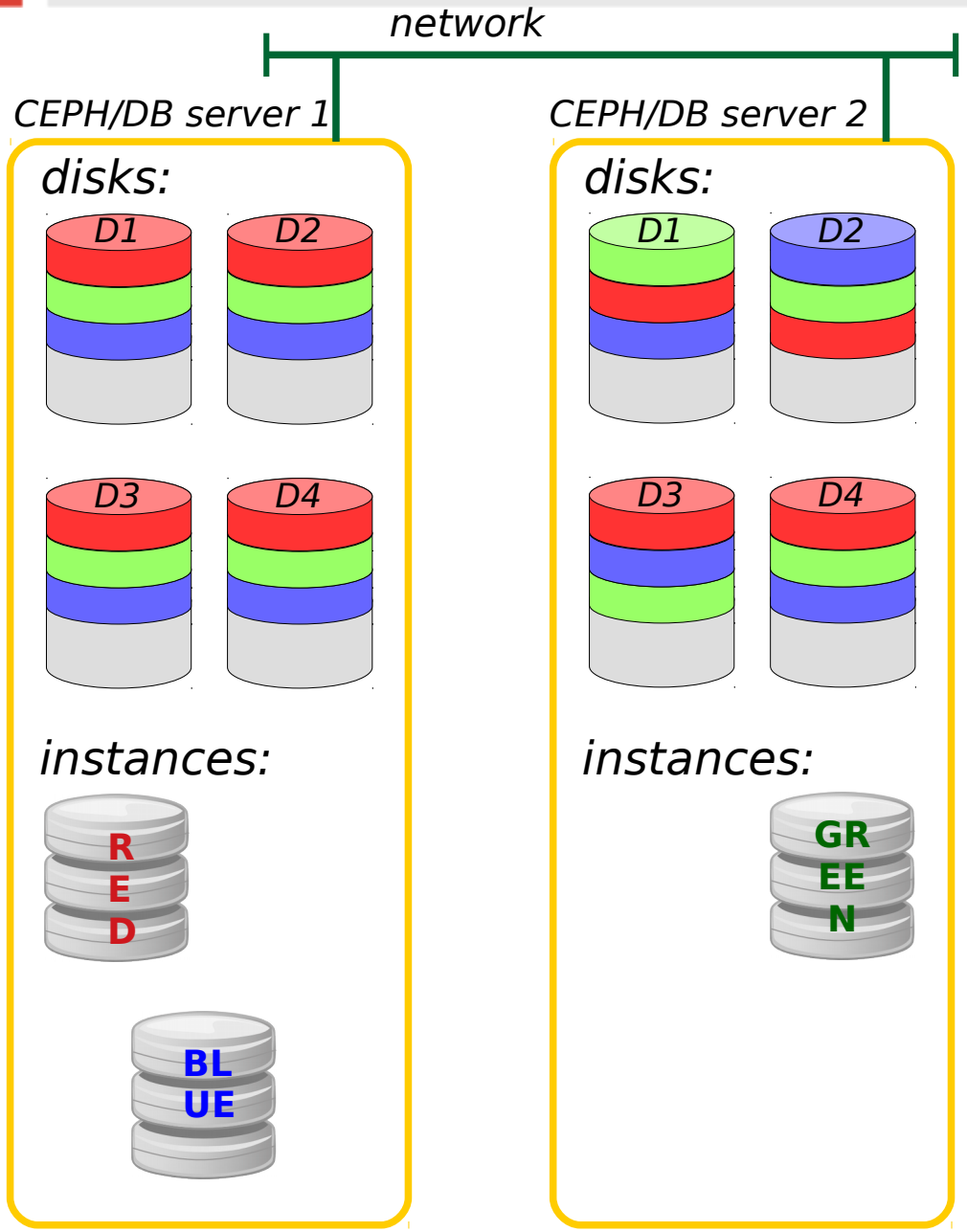
disks:



instances:

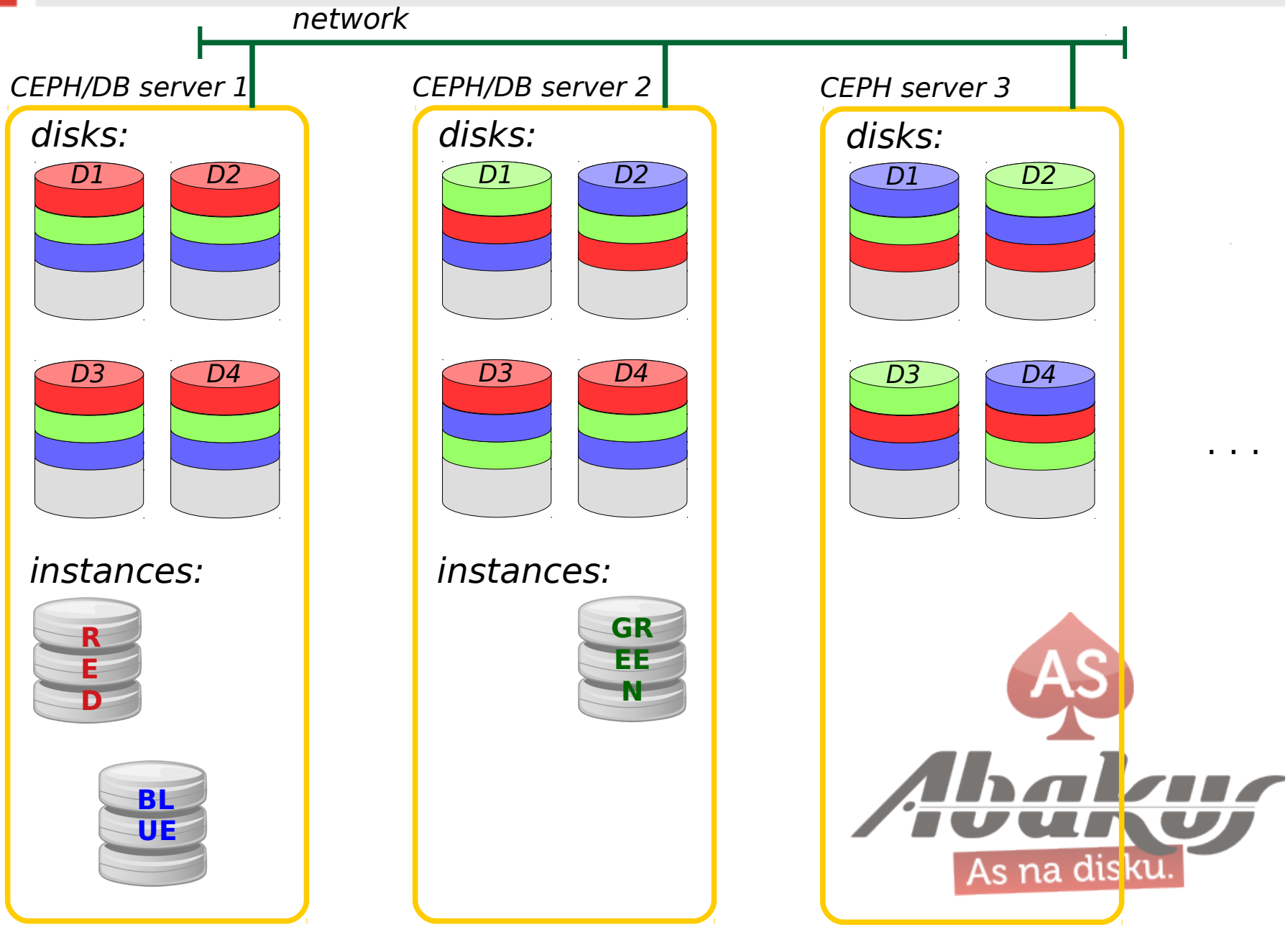


CEPH Usage – Oracle DB Server



Abakus
As na disku.

CEPH Usage – Oracle DB Server



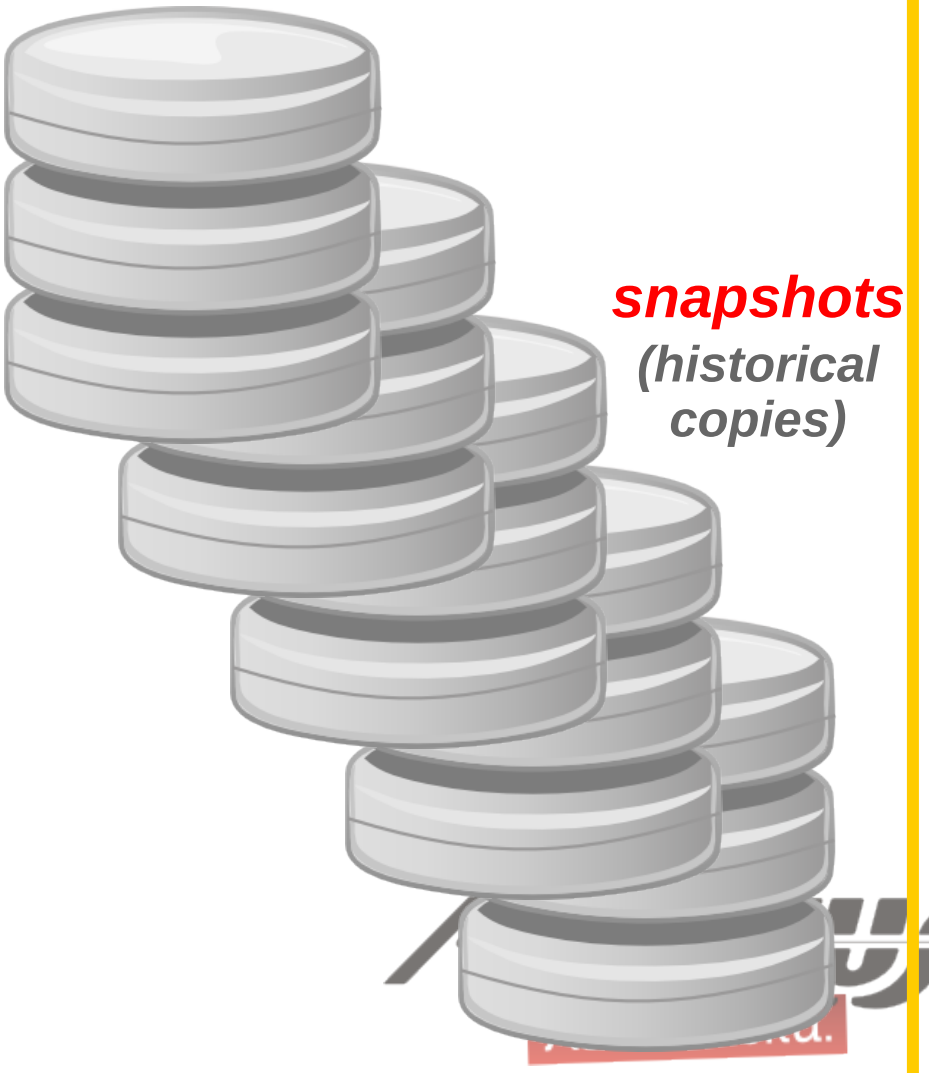
CEPH Usage – Backup Server

CEPH BACKUP SERVER

production database on CEPH

CEPH RBD replication

most recent copy



snapshots (historical copies)

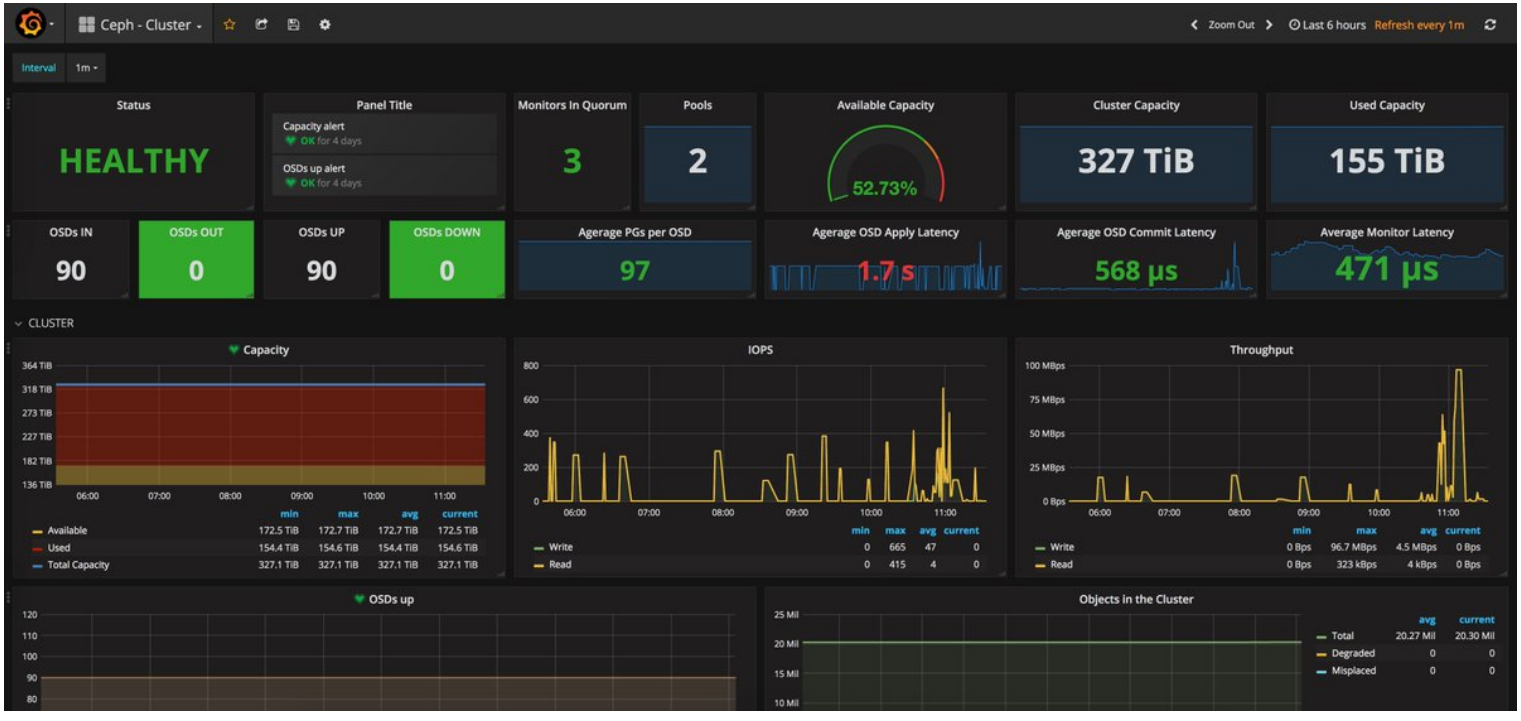
CEPH GUI

Management

- Ceph-dash
- Calamari
- Inkscope
- OpenATTIC
- VSM

Metrics

- Collectd
- Graphite
- Grafana



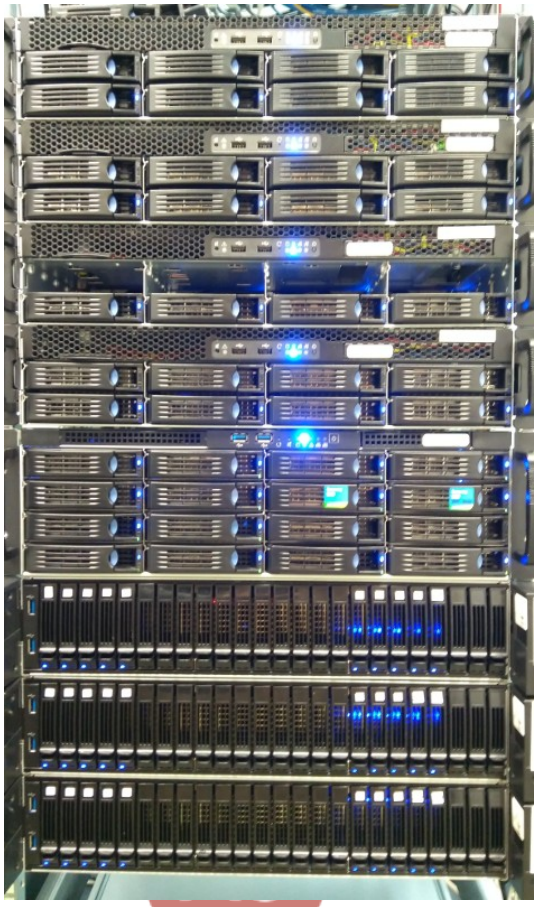
Conclusion

Mellanox blog – http://www.mellanox.com/blog/

- Ceph storage is great.
- It’s flexible – you can use it for file, block, and object storage – even at the same time.
- It’s huge – in cloud environments, containers, microservices – the modern architectures.
- It’s open – you can run it on any hardware you want.
- It scales – you can keep adding storage nodes without the need for painful data migrations.
- And it can be free – you can run the open source community version, or purchase support.

RAID is Dead, Long Live CEPH

Thank You



mag. Sergej Rožman

ABAKUS plus d.o.o.

Ljubljanska c. 24a, Kranj, Slovenija

e-mail: sergej.rozman@abakus.si

phone: +386 4 287 11 14