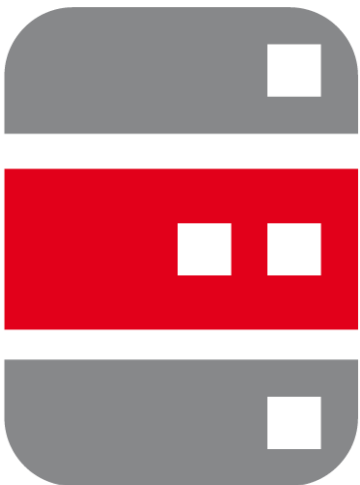


performing databases



Your reliability. Our concern.

42 (alternative?) Facts for Oracle Grid Infrastructure, ASM and RAC

Martin Klier 

Performing Databases GmbH
Mitterteich / Germany



Speaker

- Martin Klier
- Solution Architect and Database Expert
- My focus
 - Performance Optimization
 - High Availability
 - Architecture DBMS
- Linux since 1997
- Oracle Database since 2003



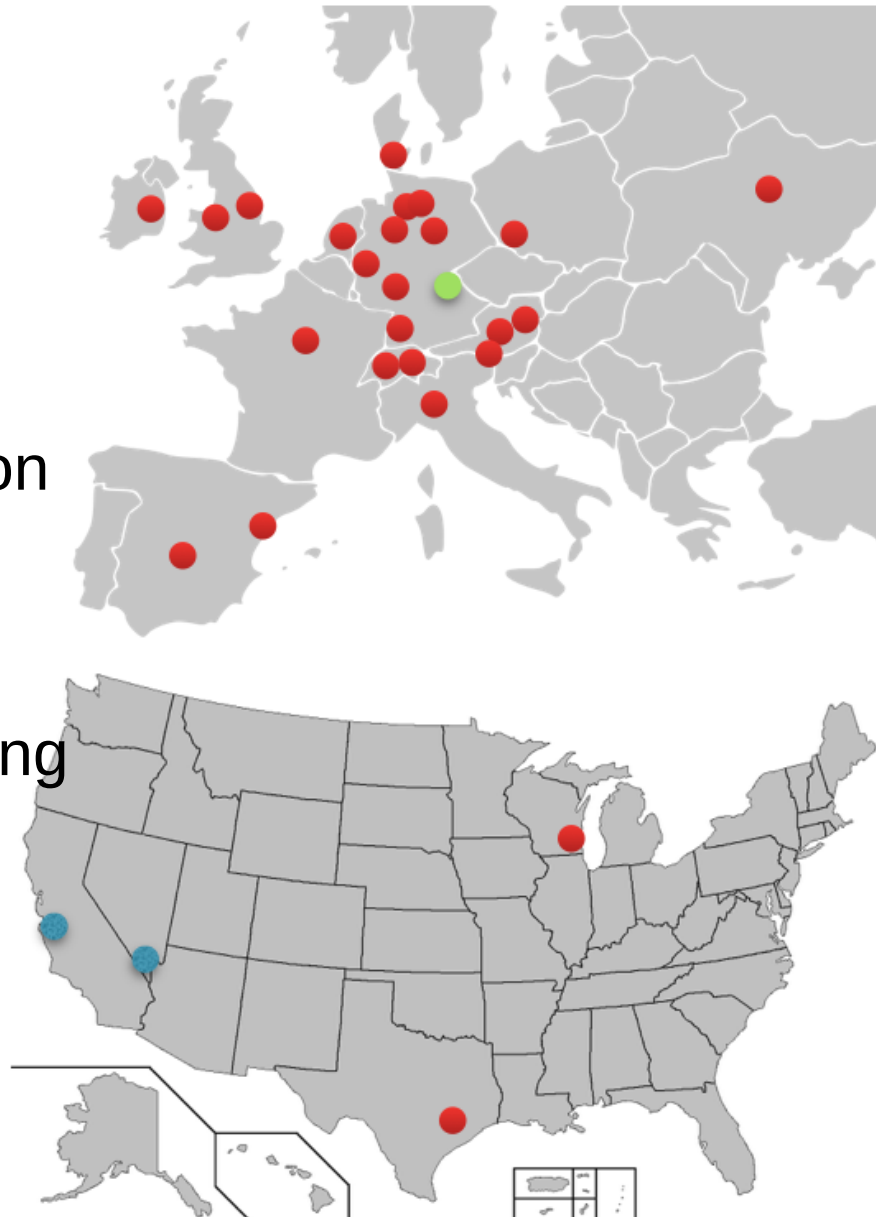
Performing Databases

- Experts for the Oracle Database
 - Concept
 - Planning & Sizing
 - Licensing
 - Implementation and Troubleshooting
- Get in touch
 - Performing Databases GmbH
Wiesauer Straße 27
95666 Mitterteich, GERMANY
 - Web: <http://www.performing-databases.com>
 - Twitter: @PerformingDB



International

- Design
- Licensing
- Implementation
- Tuning
- Troubleshooting
- Service
- Upgrade
- Migration



See you...?



DOAG



Save the date: Nov 21st - 24th 2017
2000+ Oracle Professionals
Call for papers opens soon!
<http://www.doag.org>

- Weblog: <http://www.usn-it.de> (English)
- Contact: martin.klier@performing-db.com

Tally-Ho!

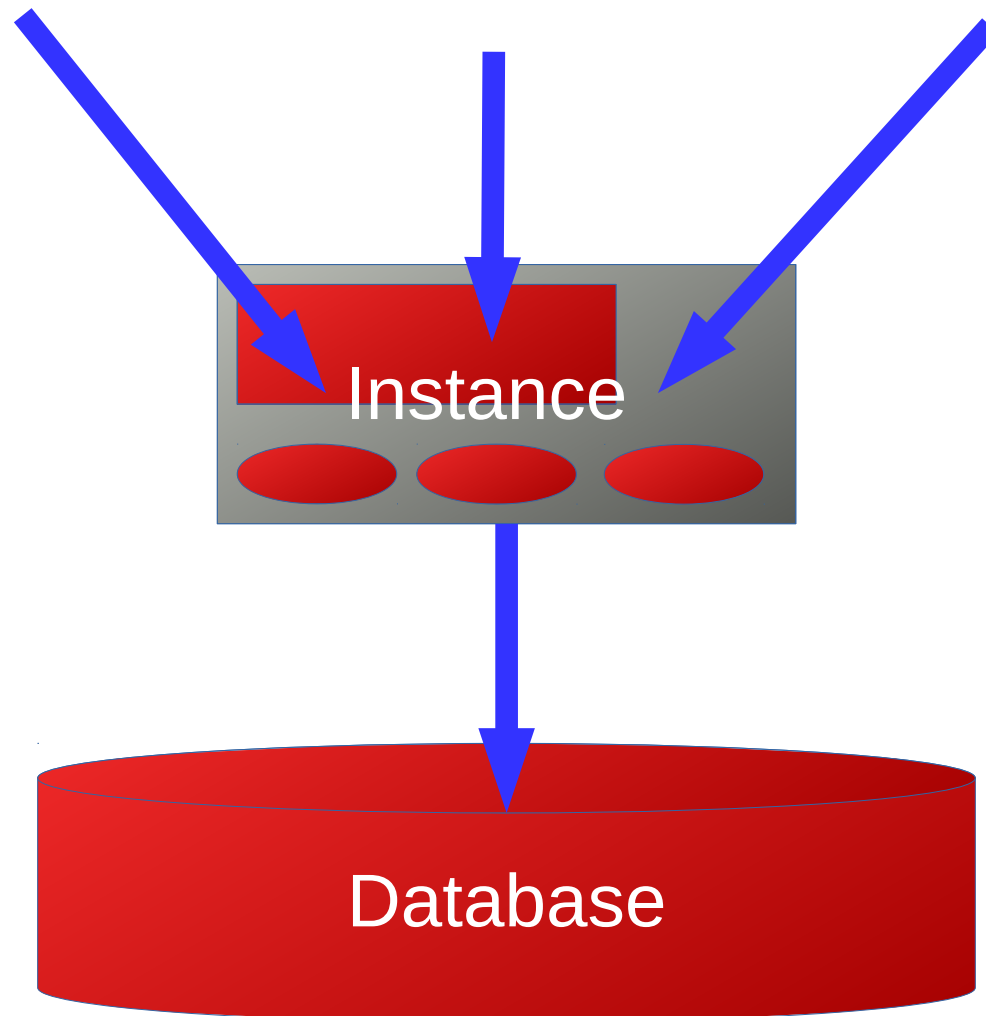


Are you heroes or wimps?

Wimps, but of the HARD kind!

Basics

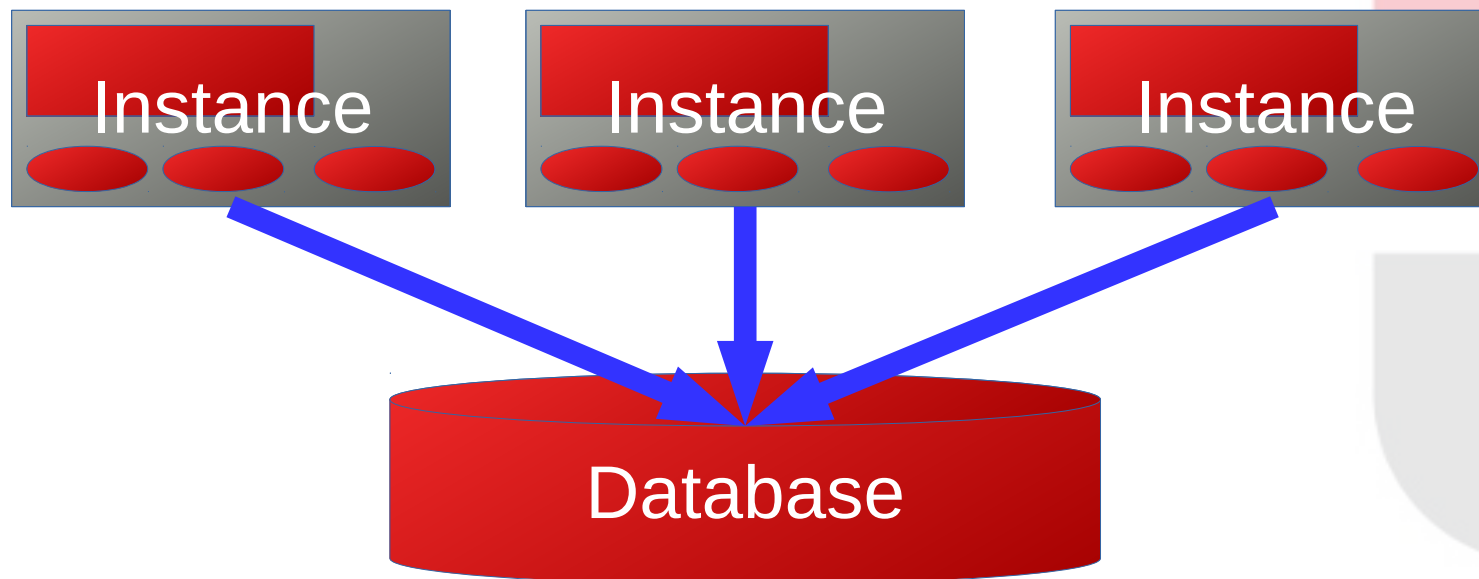
Database



RAC

RAC

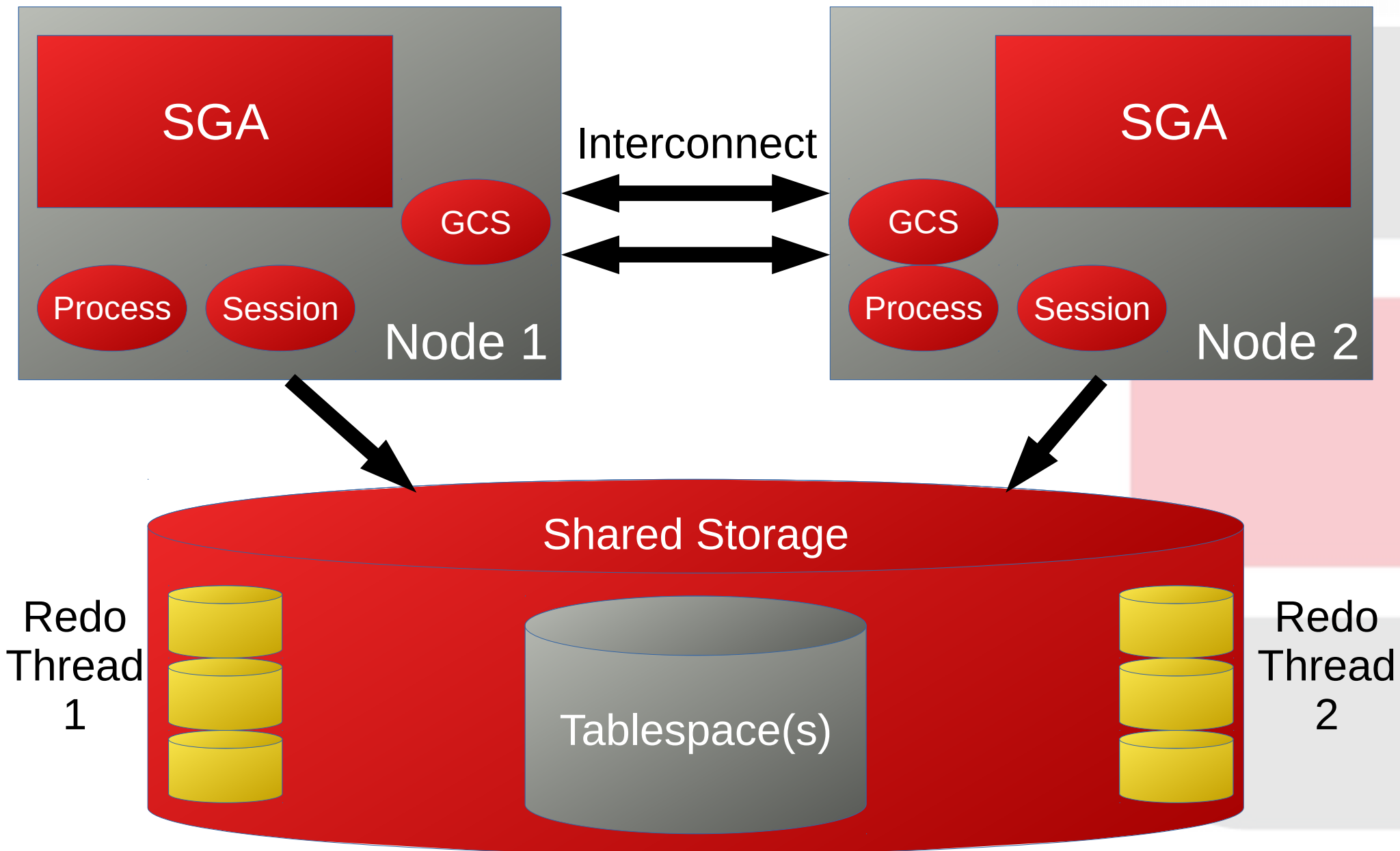
- Real Application Clusters
= Product
- Real Application Cluster
= Grid Infrastructure + Database
w/ multiple (possible) instances



RAC

- Shared-Everything-Architektur
(all nodes see all data)
- Queries will return same result on all nodes
- CacheFusion
(US Patent US20060117074 A1 by Ahmet K. Ezzat)
creates virtual global cache
(but on demand => Block Shipping)
- Parsing singleton per node (Scalability!)
Execution plans may differ - careful!

Cache Fusion



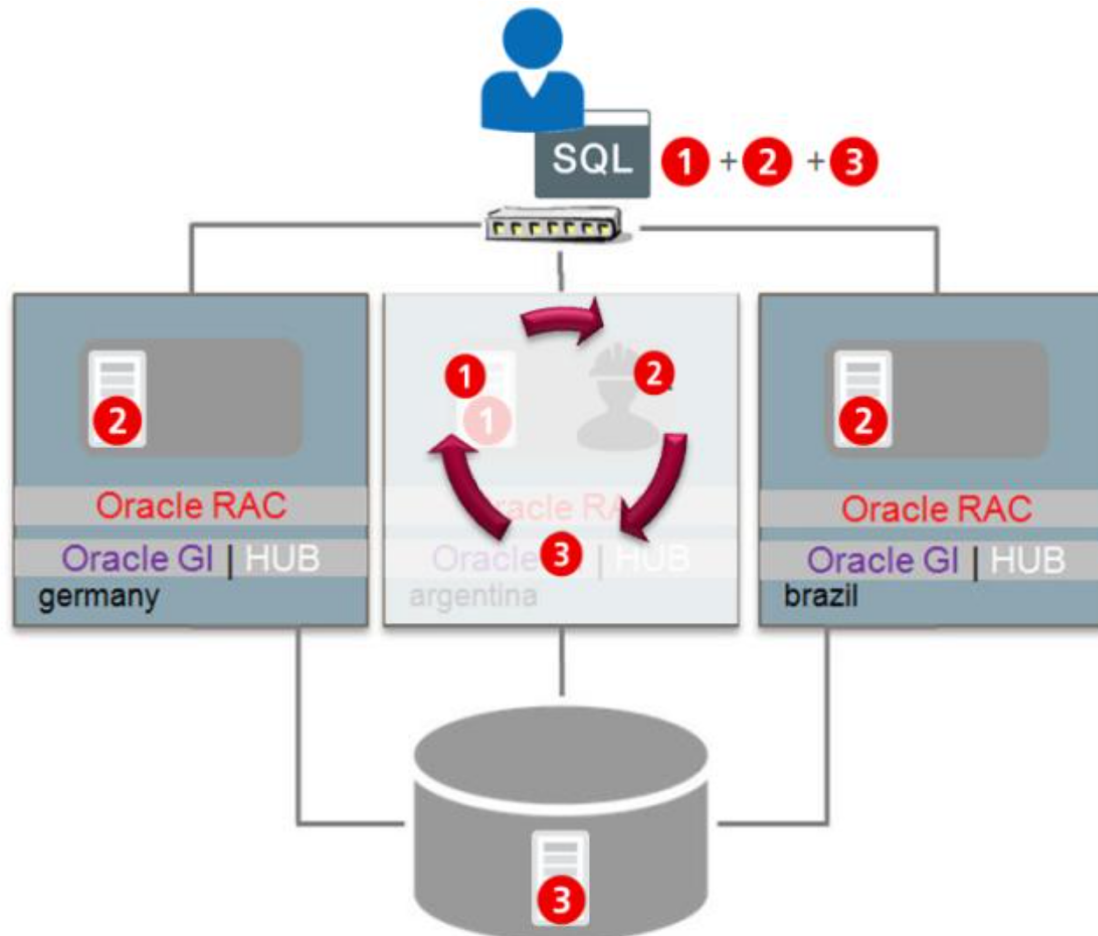
Cache Fusion

- Three ways to access a block/buffer
 - Local Cache Hit
 - Remote Cache Hit
 - Disk Access
- Global Ressource Directory
(updated by lightweight msg.)



Cache Fusion

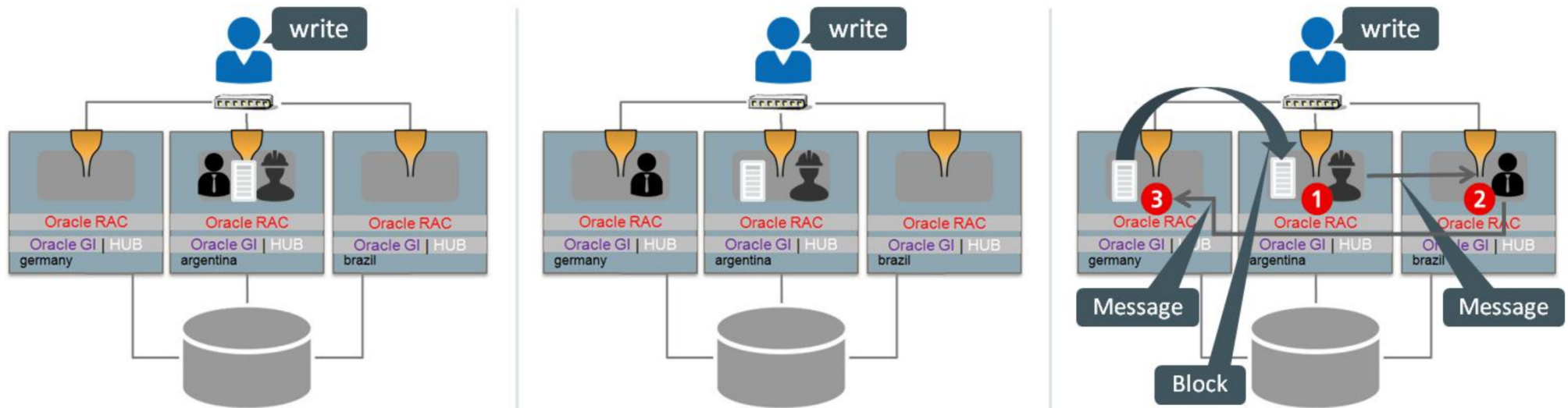
- Dynamic Data Retrieval



Picture: Courtesy of Markus Michalewicz, Oracle

Cache Fusion

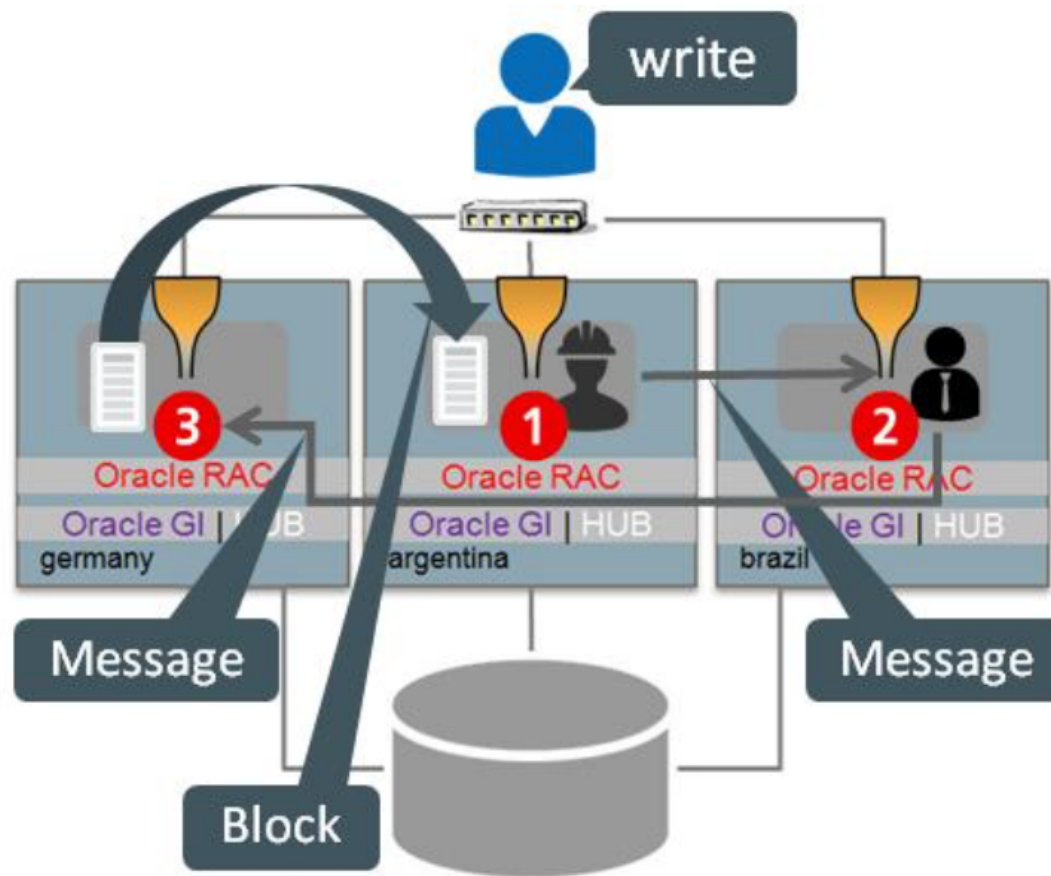
- Dynamic Mastering / Remastering



Picture: Courtesy of Markus Michalewicz, Oracle

Cache Fusion

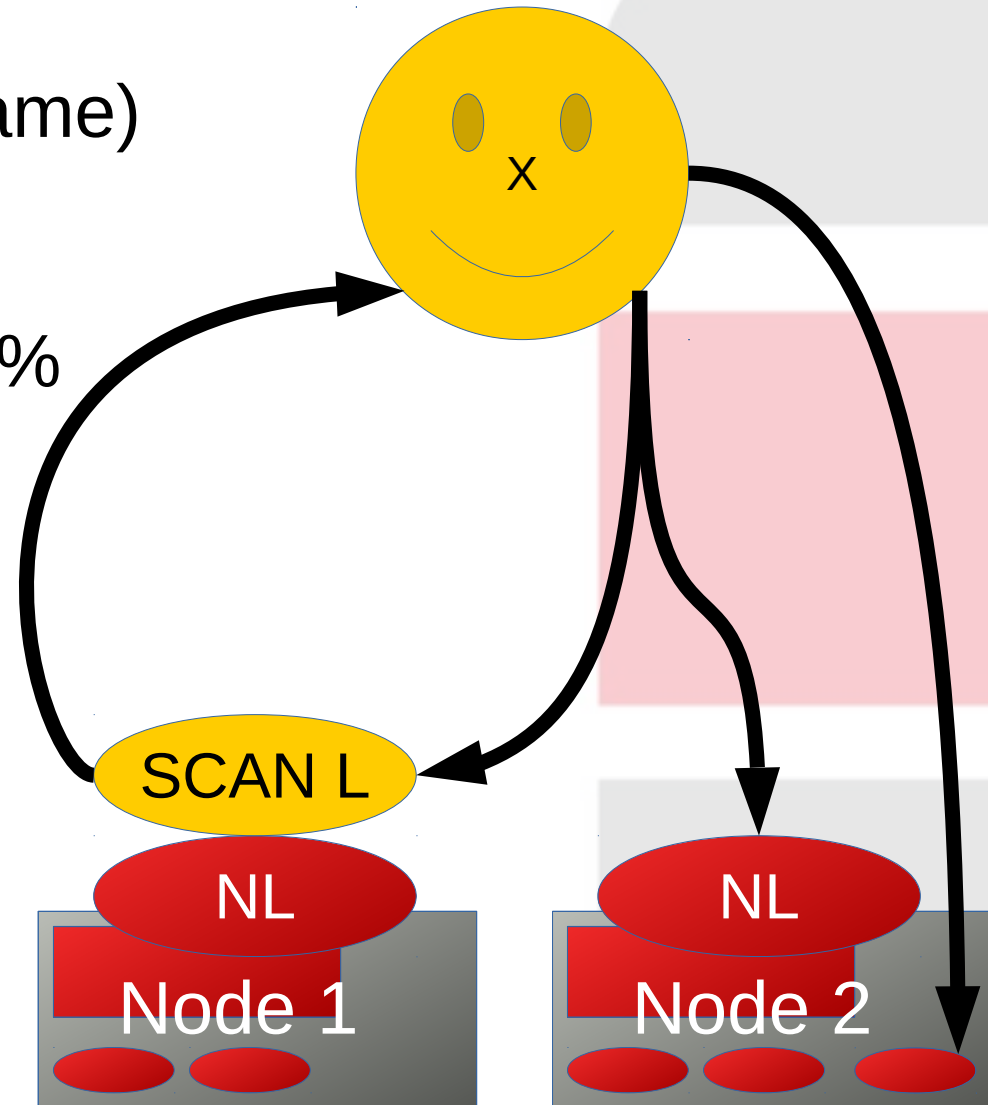
- Dynamic Mastering / Remastering



Picture: Courtesy of Markus Michalewicz, Oracle

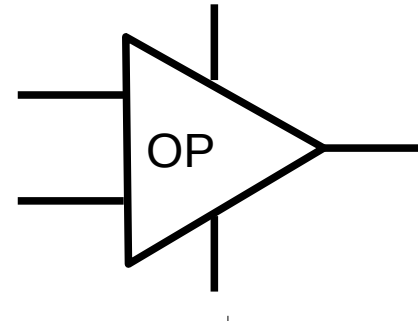
Listener

- SCAN Listener (11.2)
(Single Cluster Access Name)
- Standard: 3x SCAN
My tip: 1x sufficient for 95%
- Node-Listener

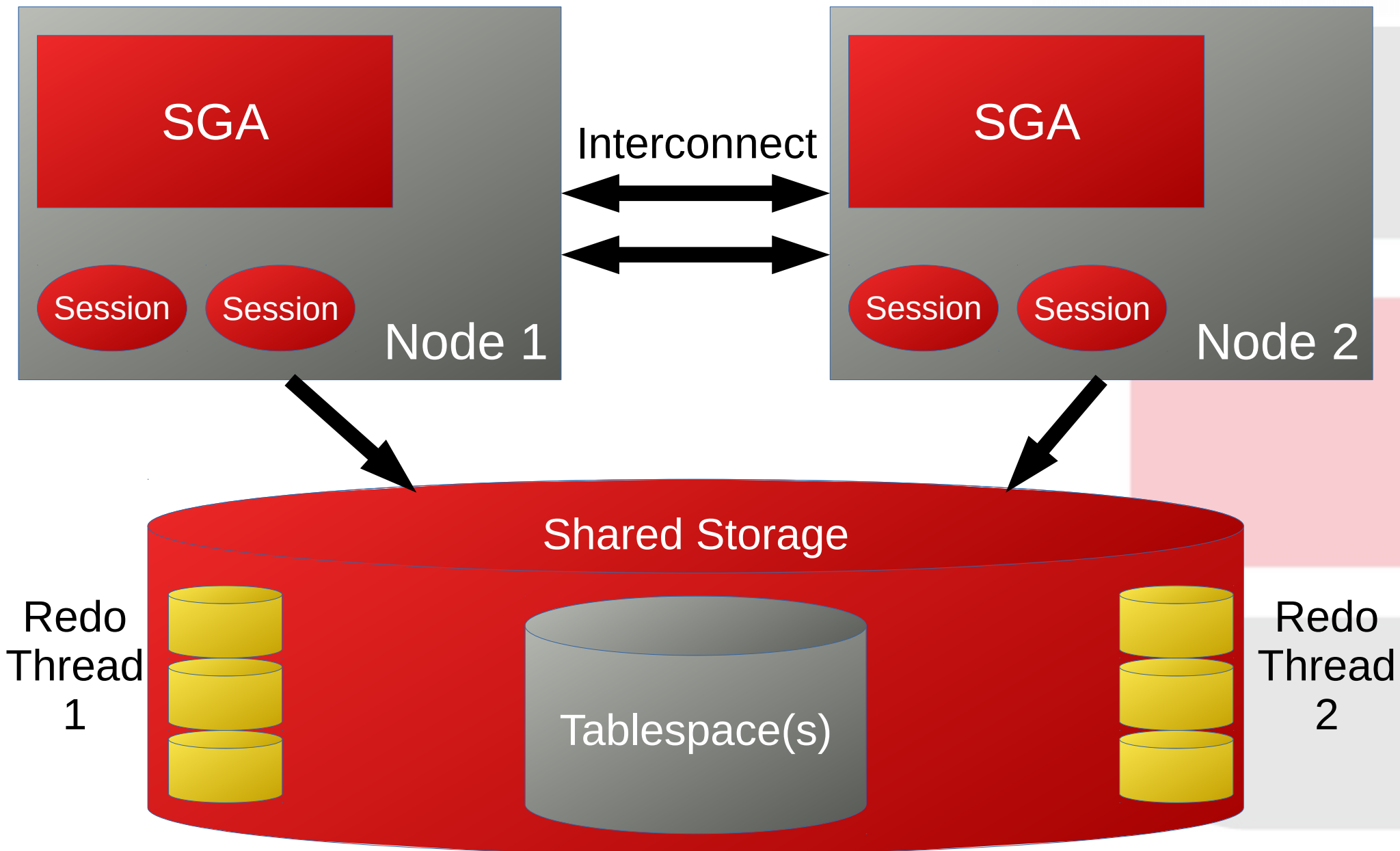


RAC Performance

- RAC = Amplifier :)
- Sequences
 - ordered
 - cached
- 1 Query not „larger“ than 1 node
- Avoid saturation (= latency degradation) of the interconnect



RAC Performance



Miscellaneous

- Auto-DOP (Degree of Parallelism) in RAC
=> Distribution PX worker over the nodes
Of interest for InMemory w/o Exa!
- InMemory IMCUs w/o Exa NOT redundant
Reason: Implemented with RDMA
=> Only works w/ Infiniband
(Mellanox requests defined environment)

ASM

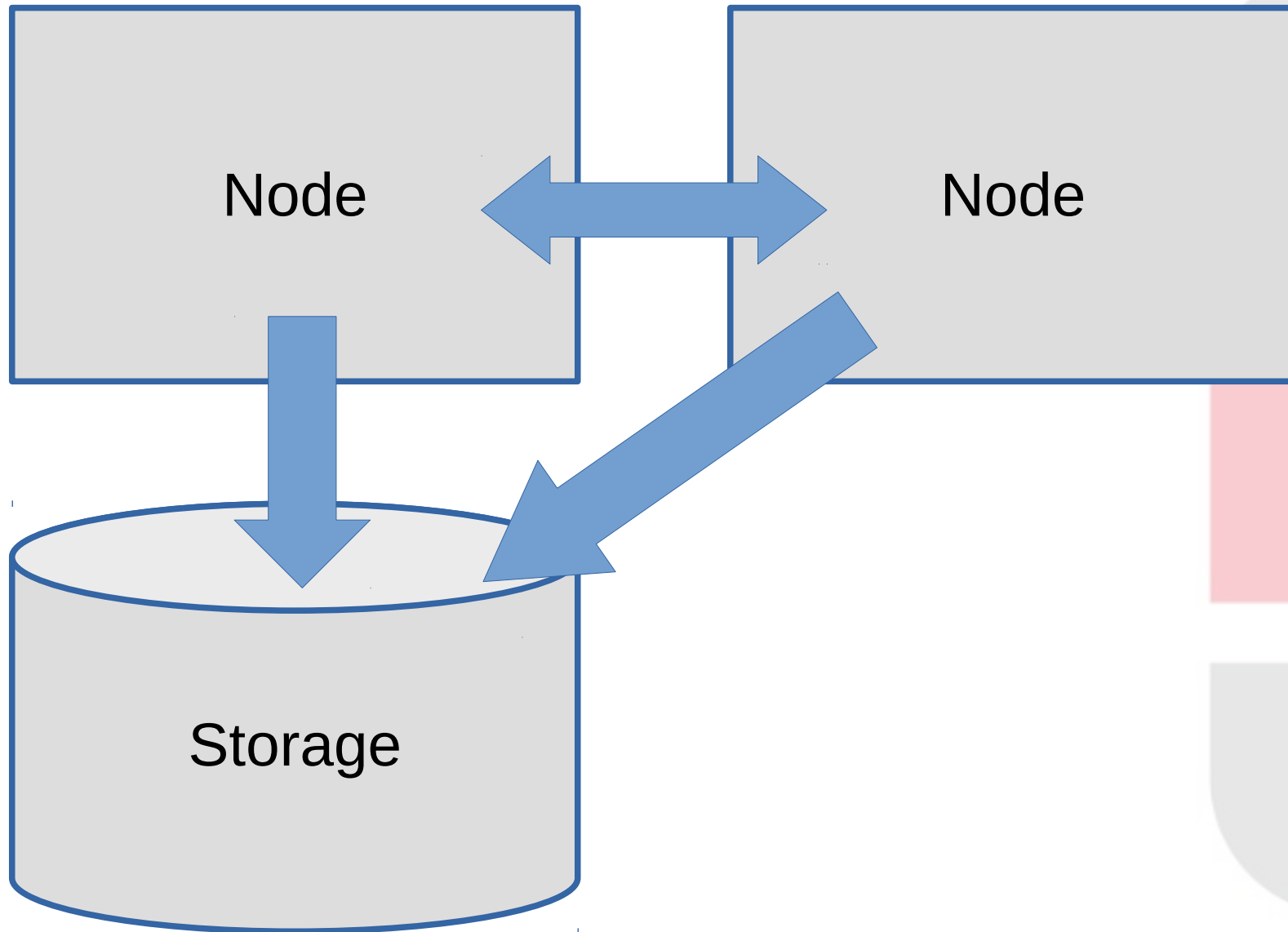
ASM

- Automatic Storage Management
= Product
- Logical Volume Manager
- Software-RAID
- Cluster - ~~Filesystem~~
- **Cluster - LVM**

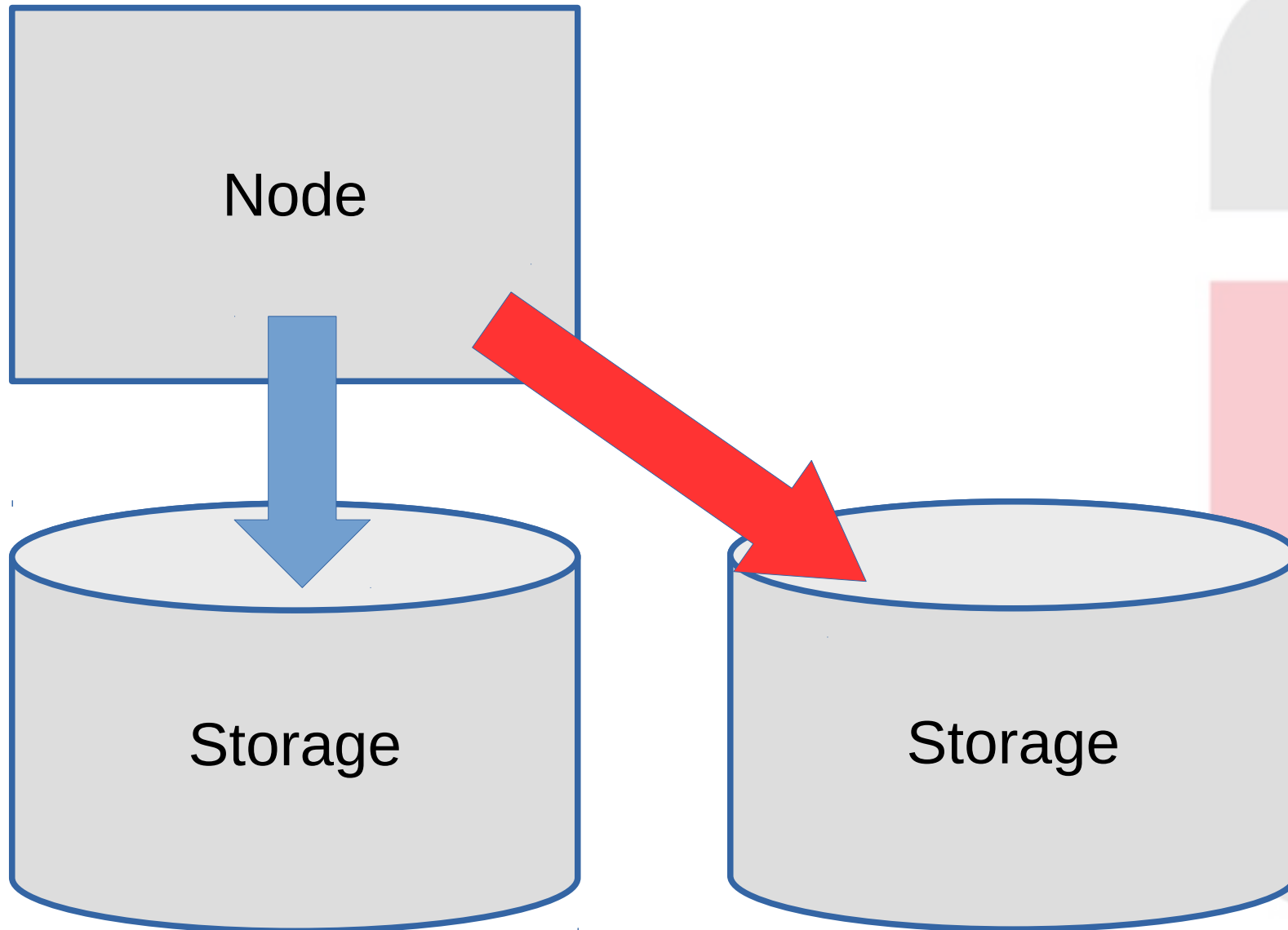
ASM

- Combines advantages of Raw Devices (small overhead) and file systems (simplified administration)
- DB will access block device directly
- Bad reputation: „I want my datafiles under my control!“
- „Humane“ Administration via multiple UIs (SQL*plus, asmcmd -p, EM CloudControl)

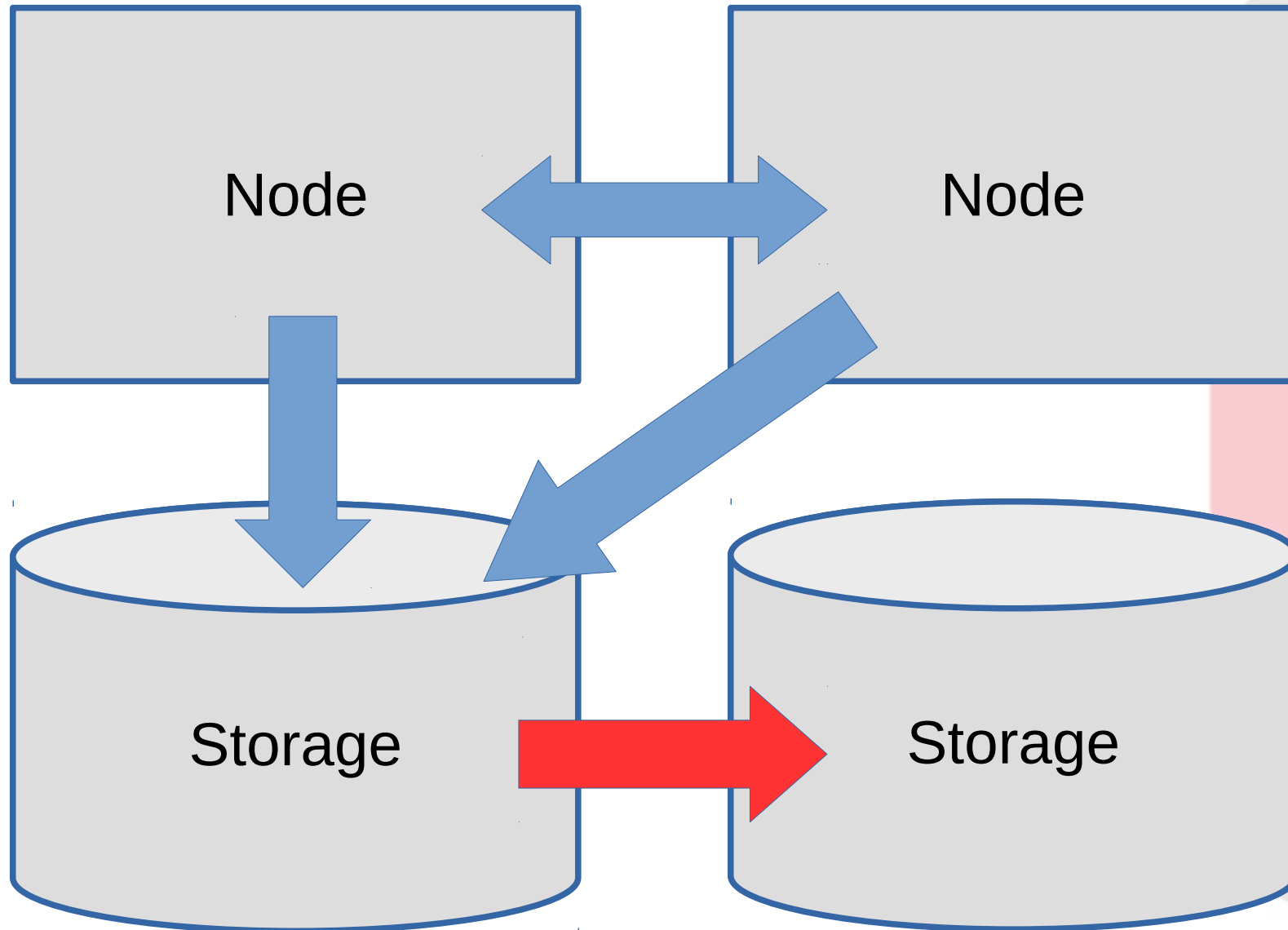
Classic I



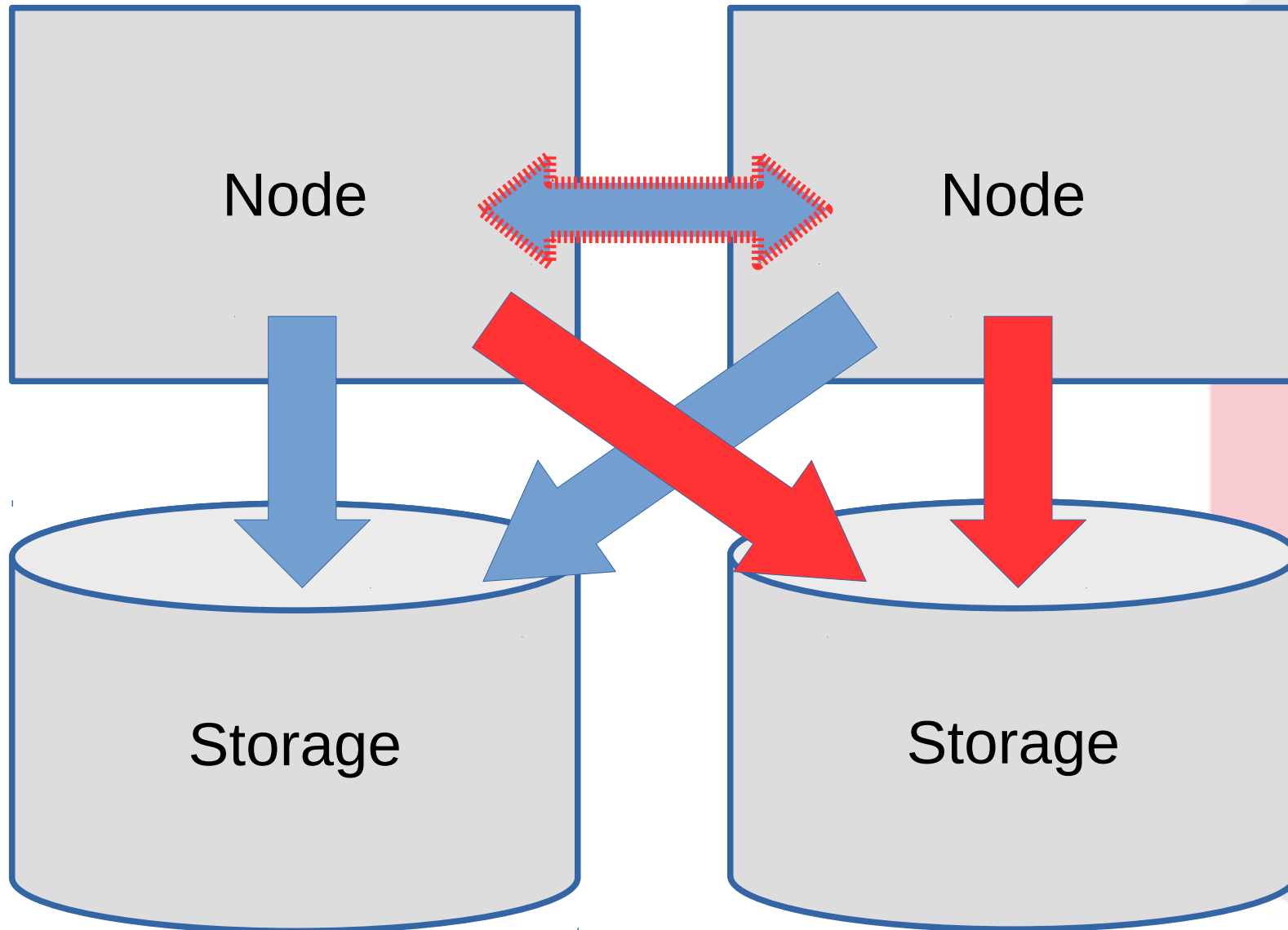
Classic II



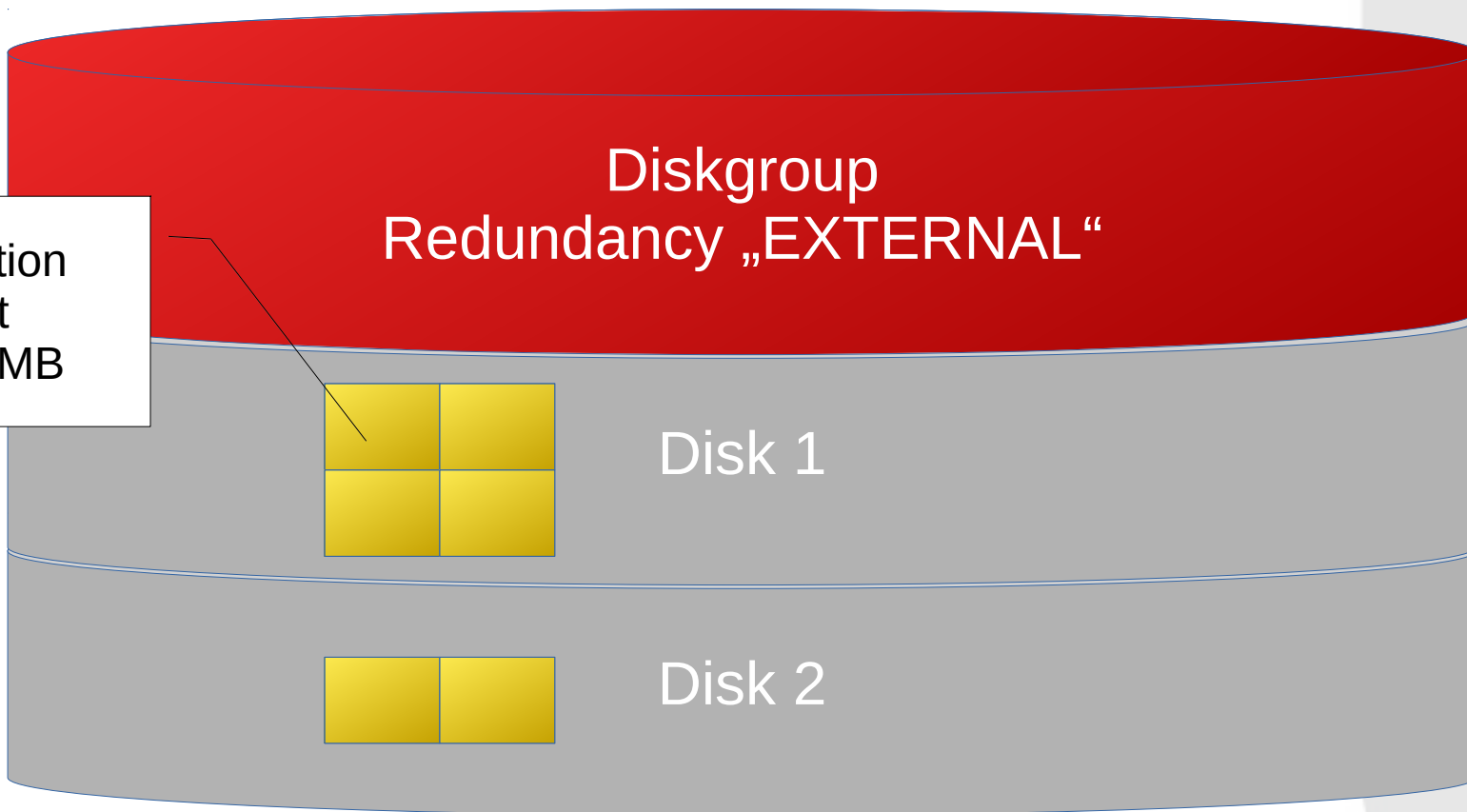
Classic III



ASM++

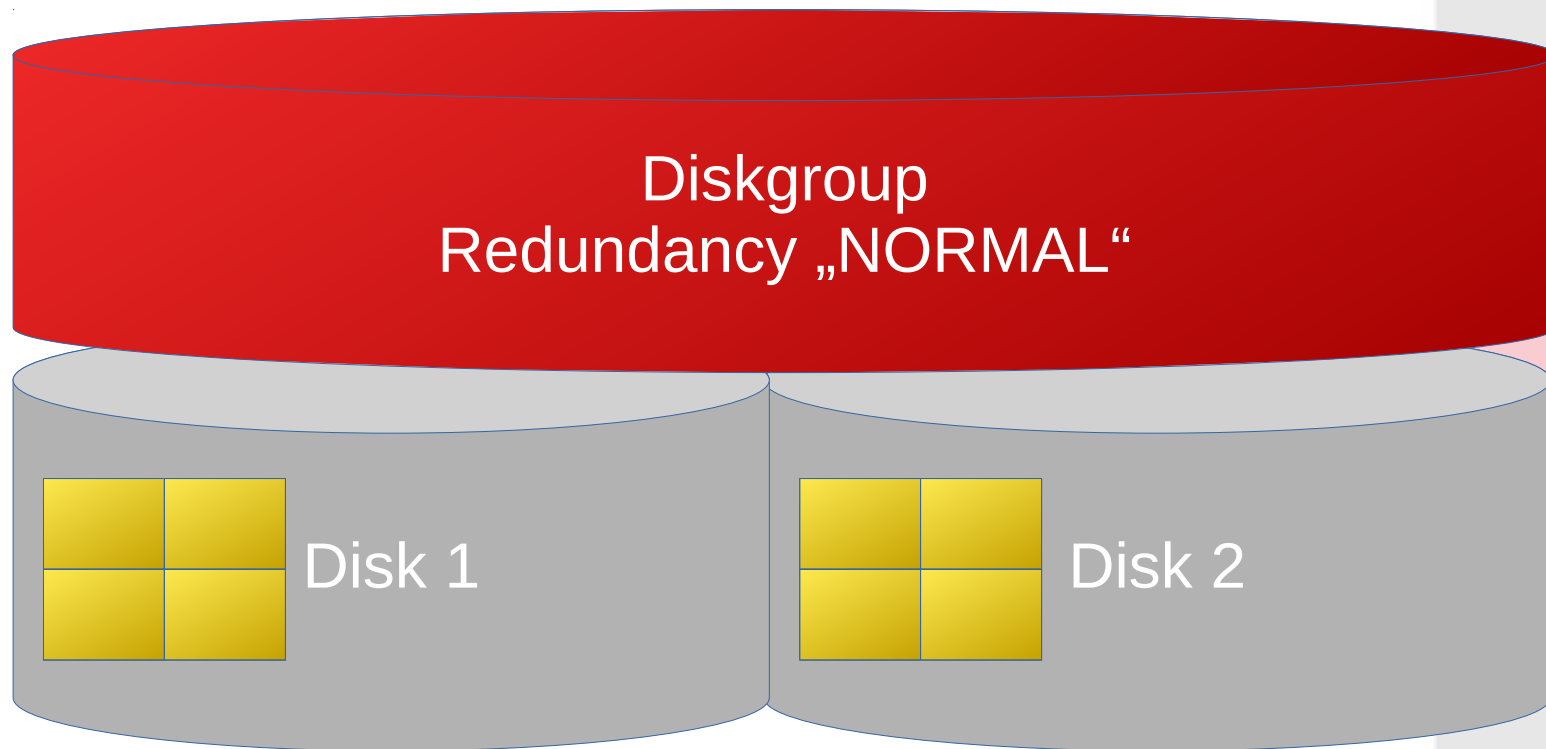


ASM Diskgroups

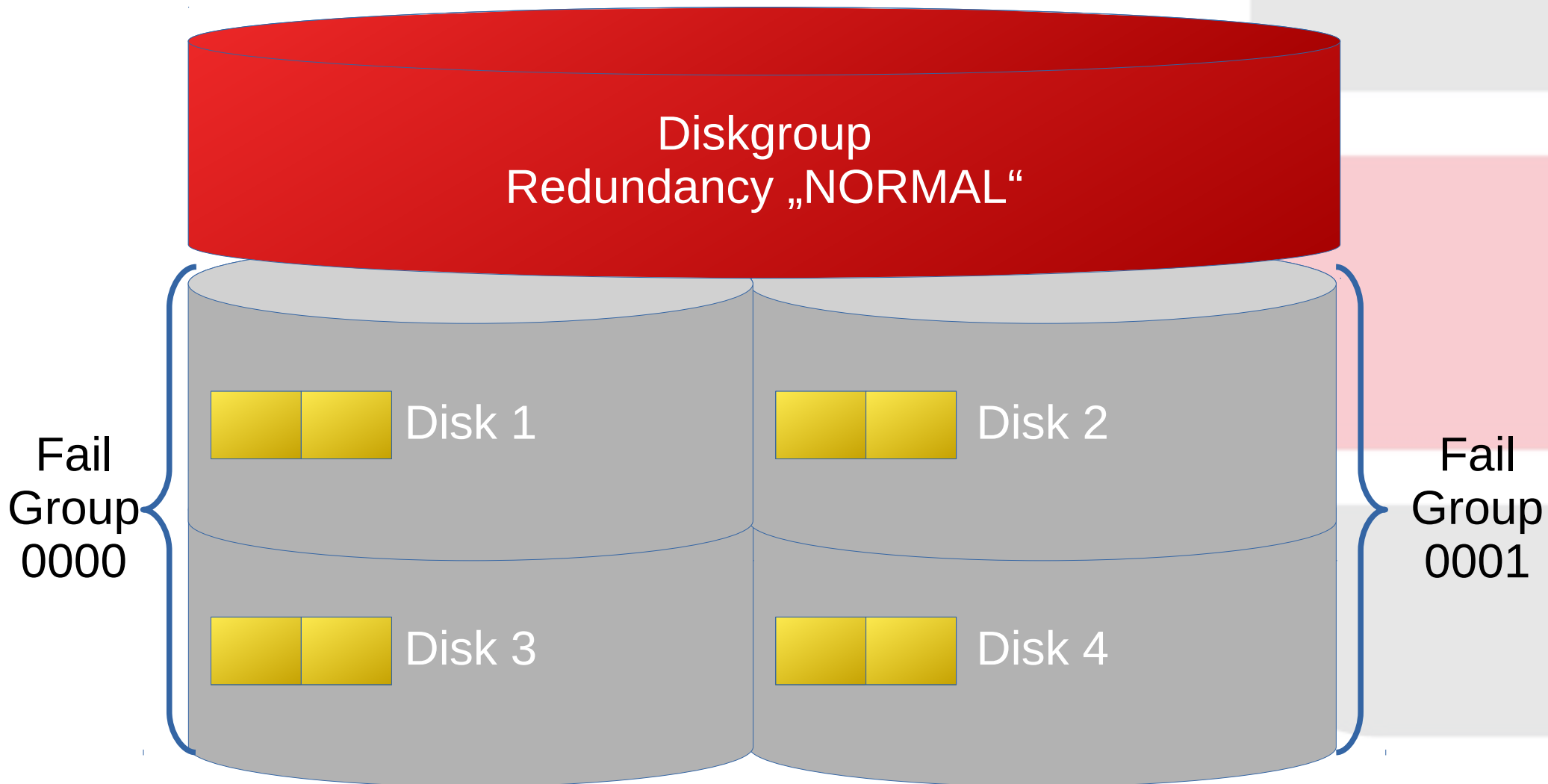


Use SAME SIZE for Disk 1 and 2!

ASM Diskgroups

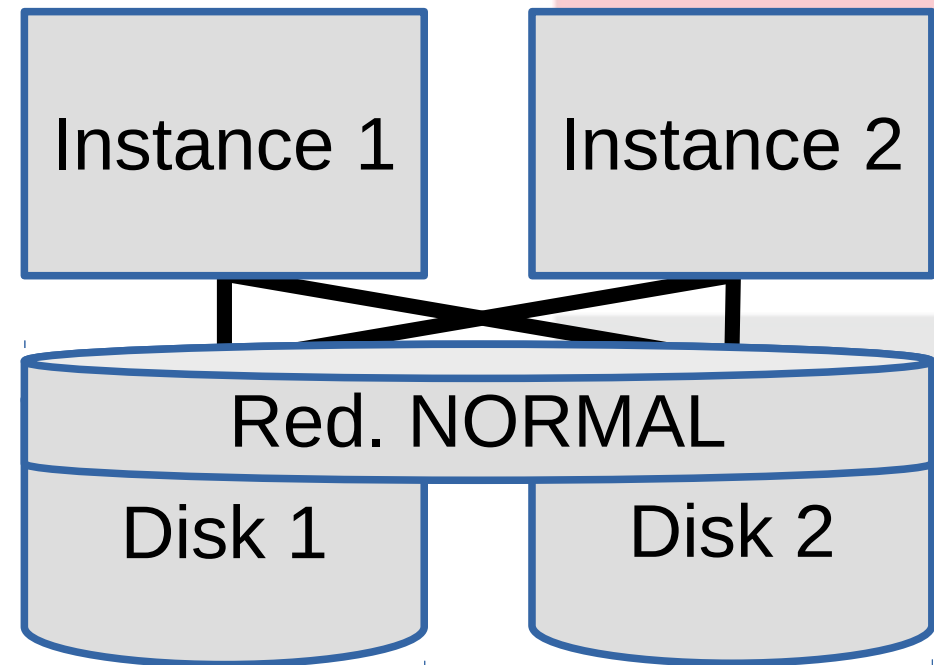


ASM Diskgroups



ASM

- Preferred Read Node (11g)
allows „local first“
- Even Reads (12c)
reads evenly from all mirrors of the DG



ASM Experience

- No data loss in >12 years
(But it was a close shave ...)
- Biggest problem in 10.x
Short-time missing LUNs not dropped + rebalanced
- „Solution“ in 11.2
 - Improved event
 - DISK_REPAIR_TIME (drop delay)
- Biggest operational challenge
Returning LUNs request manual intervention
=> Monitoring!

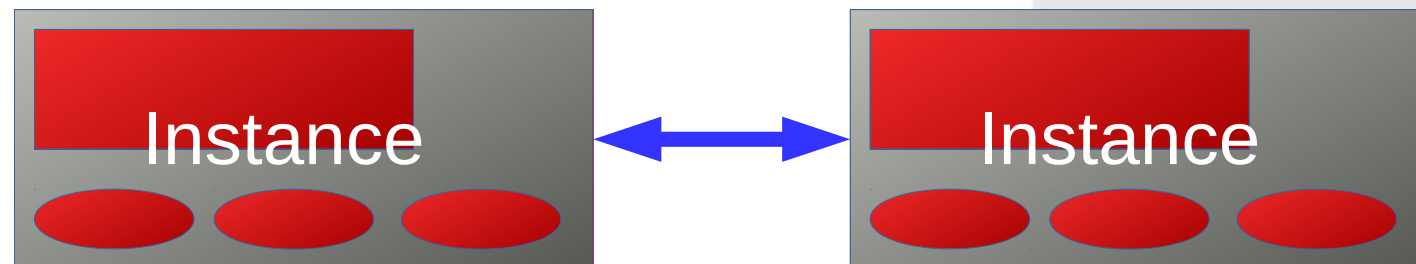
ASM 12.2

- Flex Diskgroup
 - allows grouping by database within DG
 - allows Split-Mirror backup by database/tag within DG
- ASM Filter Driver
Protection against „hostile“ IO
- ASM Service
 - centralized management of Shared Disks for all RACs in a domain
 - Central Shared Disk Service („ASM via TNS“ = NAS) will come with a patch set

Clusterware

Oracle Clusterware

- alias OCW
 - Grid Infrastructure (GI)
 - Cluster Ready Services (crs)
- Clusterware, e.g. Failover of services
- APIs and infrastructure for
 - Database (CacheFusion, ...)
 - ASM (FlexASM, ...)
 - ACFS
 - ...



OCW Services

- OHASD - Oracle High Availability Service Daemon
Starting and observing cluster processes
Interface to service mgmt of the OS
- EVM (evmd) - Event Manager
In- and outgoing messages between nodes and
services, mostly used OCW-internal
(think: Message Bus)

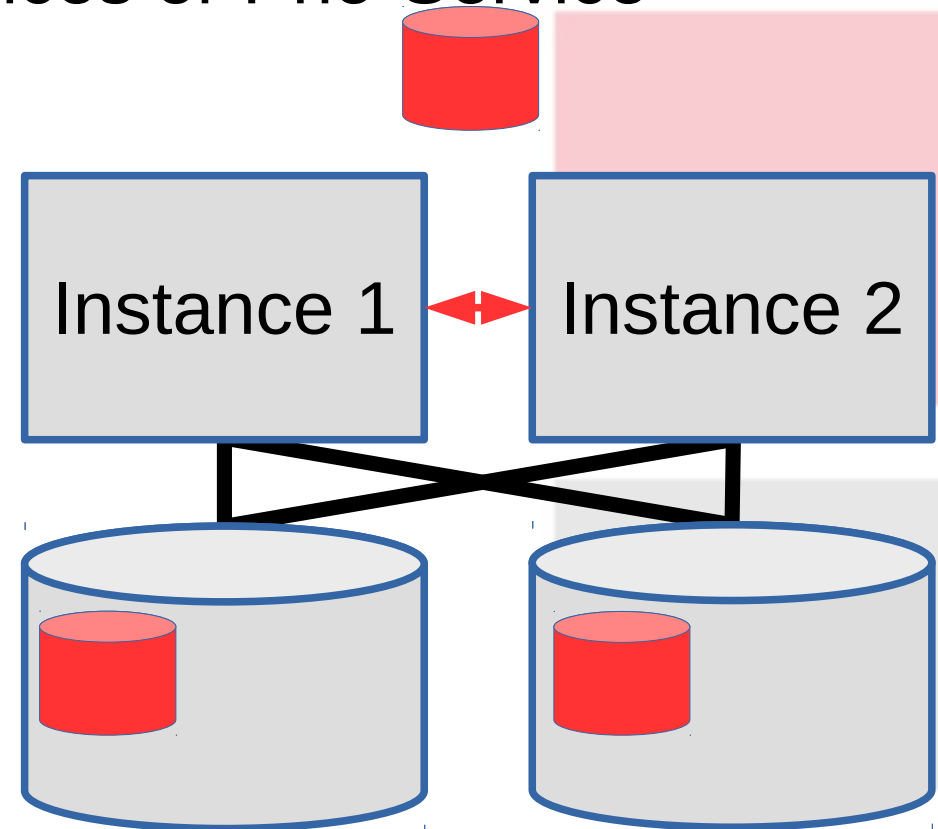
OCW Services

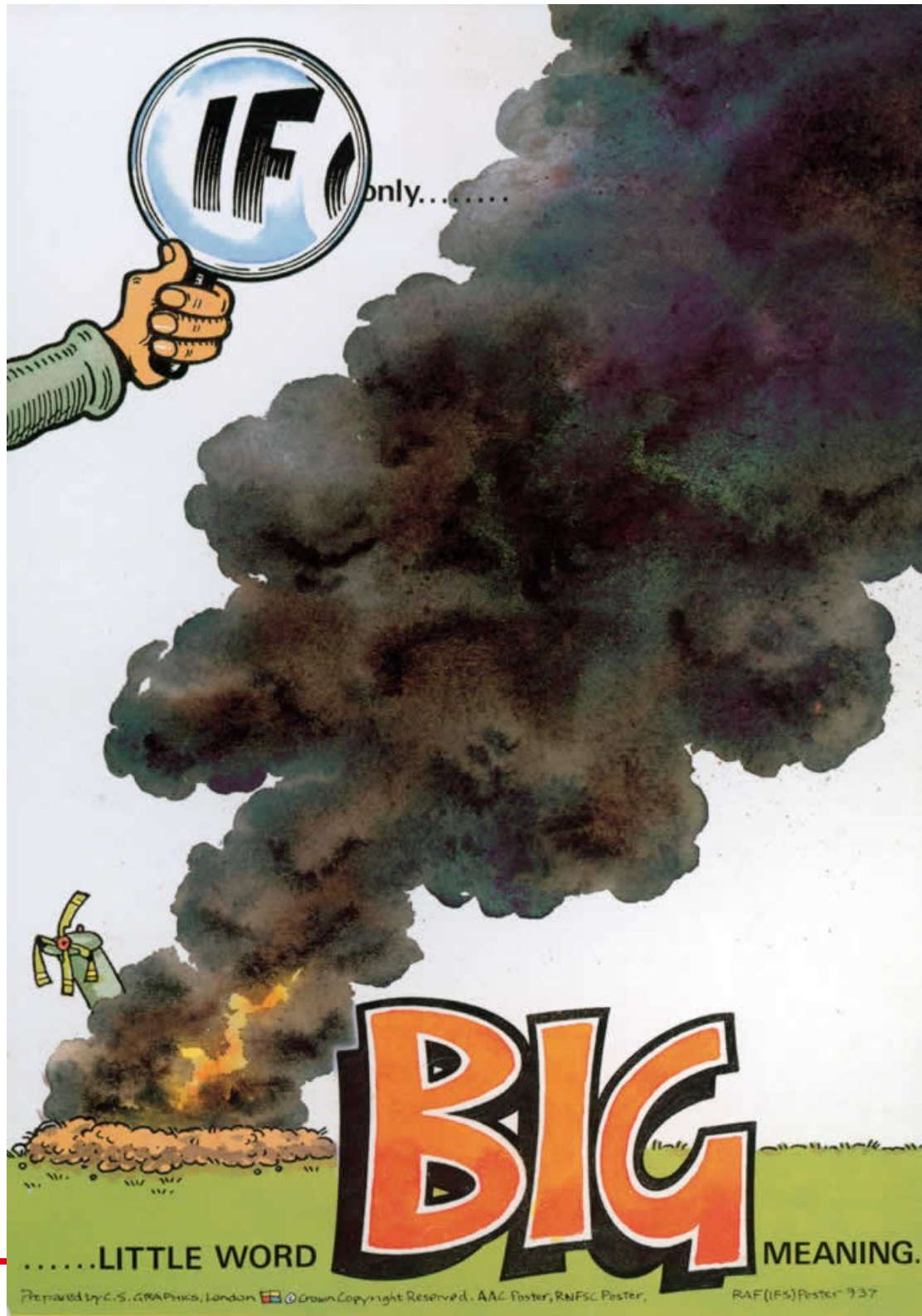
- CSS (ocssd) - Cluster Synchronization Service
Managing member nodes / availability
Disk/Network Heartbeat
Logging: ocssd.log / trc (12c)
- CRS (crsd) - Cluster Ready Service
Controlling the „Userland“ Services
(like DB, Listener)
Logging: crsd.log / trc (12c)
 - OraRootAgent - privileged services
 - OraAgent - other services
(Danger: both names also used for OHASD comp.)
 - Agent Logs show application output

H.A.

- Heartbeat Node Fencing
 - Up to 12.1: Lowest node number (or larger cluster fragment)
 - 12.2: Lowest No. of services or Prio-Service

- Split-Brain-Problem
- Quorum: 1,3,5,7 ...





Planning vs. Trouble

- Infrastructure for time sync MANDATORY (Node Evictions) - ntpd, chrony...
- if NTP, use -x / Chrony has similar option
- Last resort: CTSS not in observer mode
=> (sync Node <-> Node)
- Interconnect
 - Redundant (no Bonding needed after 11.2.0.2)
 - minimize and stabilize latency
 - use Jumbo Frames
 - Multicast
 - Isolate RACs!!

Cool & CLI

- `srvctl -h | grep xyz`
 - Do it with `srvctl !`
 - ALWAYS start utils from GRID-Home! (`$OH+$PATH`)
 - No ADD for Diskgroup Services
 - Auto create at first mount ;)
 - Full support for Dataguard (w/ DG Broker)
- `crsctl status resource -t`
- `crsctl -unsupported`
(e.g. to remove DB Service w/o `$OH`)

OCW & SE*

- Use Grid Infrastructure as clusterware for Failover Cluster - DB Failover w/o RAC One Node!
- So no need to link DB RAC Binaries
=> Standard Edition, SE1, SE2
=> 10-Day-Rule possible (max. 1 Storage)
- Create 1-Node-Serverpools
- Change DB Service for Single Instance to `Server_pools=*` (`crsctl -unsupported`, but works!)

Non-RAC

Did ya' know?

- \$TWO_TASK concats itself with @ to sqlplus
- Customized SQL prompt (glogin.sql) interferes with utilities (like dbca, asmca)
+SQL+> -- this prompt did cost me dearly
- Never end your ORACLE_HOME with /
(hash compare fails)
- No hyphen advised in DB_DOMAIN
(various known DB Link problems)

RMAN & Service

- RMAN restore from service name (12c)

RMAN> restore datafile 5 from service STANDBY;

- Works well with Datafiles, Control Files
- Does NOT with Archived Redo Logs :(

Environment

- Oracle says, use different OS User Grid vs. DB
- My tip: Avoid if possible.
Write (or copy&understand) very good env scripts
- Suggestion: /home/oracle/bin
- Set environment and shell prompt
like [oracle@myhost ~] (MYSID) \$
- Source with logon (.yourshell_profile)
or manually: . db / . grid

Heads in and not LOOKING OUT...



**...can seriously
damage your health**

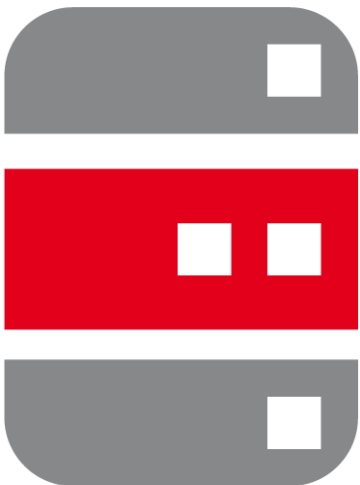
**Maintain constant LOOK-OUT at all times
However interesting the data may be!**

Defence Aviation Safety Centre Poster No 200513



Download my Presentations and Whitepapers
<http://www.performing-databases.com>

performing databases



Your reliability. Our concern.